

An auspicious victory

A vote by the US House of Representatives to ease restrictions on embryonic stem-cell research marks an important turning point — whether President Bush vetoes the change or not.

Last week, the US House of Representatives voted by 238 to 194 to reverse the policy restricting embryonic stem-cell research that was implemented by President George W. Bush back in 2001 (see page 544).

The measure will now be taken up by the Senate, where it has a good chance of success. However, Bush said both before and after the House vote that he will veto the measure if it reaches his desk. Even that won't erase the significance of this first victory in the Republican-controlled House for advocates of stem-cell research. The fact that 50 Republican members voted for the change in policy underlines the fact that the pendulum of public opinion is swinging strongly in favour of allowing more of the research to go ahead.

That augers well for the eventual loosening of federal policies that have kept scientists' hands tied in the United States. Some influential Republicans — including Nancy Reagan, the wife of a former president, and conservative Utah senator Orin Hatch — have spoken out in favour of embryonic stem-cell research, making it easier for others to publicly support the work as well.

Indeed, the desperate language used by opponents of embryonic stem-cell research suggests they know they are losing the debate. Tom DeLay (Republican, Texas), majority leader in the House, said on Tuesday last week that a vote for the stem-cell bill was a vote “to fund with taxpayer dollars the dismemberment of living distinct human beings for the purposes of medical experimentation”.

There's no indication that the public is buying this. Polling evidence suggests that most Americans instead have high hopes for the research, and biomedical advocacy groups have done an effective job convincing lawmakers that the research deserves a chance to fulfil these hopes. Years of tireless lobbying by these groups played a major role in the 24 May victory. So too did competitive pressure from

overseas, with South Korean research making headline news just ahead of the vote, and from state governments, whose own initiatives in this sphere are forging ahead in California, New Jersey and elsewhere.

Even if the measure passed by the House becomes law, the United States would retain more restrictions on publicly funded research than do nations such as South Korea and Britain, both of which allow publicly supported scientists to use somatic cell nuclear transfer to derive fresh embryonic stem-cell lines. The US measure would allow federal funding only for work on embryos left over from *in vitro* fertilization clinics.

In the Senate, Sam Brownback (Republican, Kansas) has already said that he will attempt to block a vote on the stem-cell measure, so that for the measure to pass it will need the support of 60 out of the 100 senators. If it obtains these votes, the bill will arrive on President Bush's desk — and he'll face a difficult choice.

“The desperate language used by opponents of embryonic stem-cell research suggests they know they are losing the debate.”

The proponents of stem-cell research have mooted the idea of some sort of compromise, which might, for example, update the 2001 policy to allow publicly funded work on more recently derived stem cells. But the president has chosen to draw what he sees as a moral line in the sand that he feels he cannot cross. He has threatened to make this the first bill that he has vetoed in four-and-a-half years in office. But the veto, if it is used, will place him squarely against public opinion in the United States, which increasingly views embryonic stem-cell research not with fear, but with hope. ■

Seeds in threatened soil

US hostility towards Syria is undermining the stability of an important seed bank for dry areas.

Thirty kilometres from Aleppo in Syria, not far from the birthplace of agriculture, is the International Center for Agricultural Research in the Dry Areas (ICARDA). It includes an international gene bank that holds seeds in trust on behalf of the world's dry countries.

Organized through the World Bank and funded by international donors, ICARDA's gene bank holds samples of 131,000 individual seeds for plants that form part of the diets of one billion people who live in Central and West Asia, the Middle East and North Africa. The seeds include different varieties of barley, beans, chickpeas and

lentils, catalogued and stored in sealed plastic bottles inside giant refrigerated vaults.

ICARDA often finds itself having to rebuild agriculture at the end of military and civil conflicts. The centre is in effect a lender of last resort for farmers and scientists who have nowhere else to go when their seeds run out.

When Taliban fighters looted Afghanistan's national seed store in 2002, they took the empty plastic bottles, leaving the seeds behind. Even so, the country's scientists needed ICARDA's help to rebuild the store. And shortly before the US invasion of Iraq in March 2003, Iraqi scientists sent a 'black box' across the border to ICARDA containing copies of the country's seed stocks. The action was timely, as Iraq's seed bank, in the Baghdad suburb of Abu Ghraib, was looted and destroyed during the insurgency. ICARDA plans to use the contents of the box to help regenerate Iraqi farming.

But now the centre's host country is itself feeling the heat of US



rhetoric. The US government has always been a generous financial supporter of the centre's activities. But in the words of the State Department, Syria is autocratic, is a state sponsor of terrorism, and is believed to be developing weapons of mass destruction. Continuing US sanctions and some discussion in the United States about

"\$260 million is a relatively small price to pay to conserve the world's food heritage and secure the future supply of food for the world."

possible 'regime change' are causing nervousness.

One response would be to pack the seeds into storage boxes and airlift them out of Syria, but the threat of US military action currently seems too remote to warrant such drastic action.

Much better, for ICARDA and for the 14 other 'Future Harvest Centres', would be for more support to be given to the Global Crop Diversity Trust, an international fund to build more gene banks around the world and to improve the conditions of existing ones. The trust was set up jointly by the United Nations' Food and Agriculture Organization (FAO) and the Consultative Group on International Agricultural Research. It says it needs an endowment of \$260 million to safeguard seeds used in world agriculture and to improve the condition of the gene banks where they are stored.

The world's gene banks are in a parlous state, as a new report ("Safeguarding the future of US agriculture") published jointly by the US Department of Agriculture and the University of California makes clear. Of the 1,460 gene banks around the world, only 35 meet international standards for long-term storage. These include the gene banks of ICARDA and of the other Future Harvest Centres. The FAO, moreover, says that nearly-one fifth of the 5.4 million seeds stored in gene banks are degenerating.

The US report also urges the Bush administration to support the Global Crop Diversity Trust, and not without good reason. Pests and plant diseases are causing losses to US agriculture of up to \$33 billion each year, and there is a strong fear that new threats could cause even more damage. US agricultural researchers are currently scouring the world's gene banks for seed varieties that can resist these diseases. Chief among such diseases are a fungus that is currently invading US soybean fields, and potato blight of the kind that caused the Irish potato famine, which is destroying potatoes worth some \$400 million each year.

The US government is currently spending more than \$1 billion per week on military operations in Iraq. By comparison, a \$260-million endowment is a small price to pay to conserve the world's agricultural heritage and to secure the future food supply of the United States and the rest of the world. ■

Too much, too soon

How not to promote your latest research findings in the media.

A tour de force; an impressive advance; years ahead of its time. When South Korean researchers declared last month that they had created stem-cell lines genetically matched to individual patients, commentators were ready with superlatives, and rightly so. The paper (W. S. Hwang *et al.* *Science* doi:10.1126/science.1112286; 2005) is a major step towards the use of stem cells in the study and treatment of disease.

In Britain, however, many usually well-informed members of the public may be labouring under the illusion that it is their nation, and not South Korea, that is pushing back the boundaries of stem-cell research. For just as Hwang's paper appeared in press, a second one — or, at least, an abstract of one — sprang forth from a team of biologists at the University of Newcastle upon Tyne. This group, led by fertility specialist Alison Murdoch, had not matched Hwang's achievements — they merely described the creation of a cloned embryo, not the extraction of cell lines — but they stole most of his thunder in the UK press.

If this were just a routine case of domestic media favouring local achievements, it wouldn't matter much. But the manner in which the Newcastle team made its discovery public has consequences that reach beyond one day's headlines. As researchers in the field have been angrily informing *Nature* since the two pieces of work appeared, the approach taken in this case risks damaging science and its public perception.

The Newcastle team submitted its work to an independent Cam-

bridge-based journal, *Reproductive BioMedicine Online*, which has the unusual policy of making abstracts of submitted papers available on its website as soon as the articles are sent out for peer review. The full paper is kept confidential until it is accepted and published. So science reporters informed of the findings by a telephone briefing had access to an abstract that had not been peer reviewed — and to nothing else.

It can't yet be determined for certain if the Newcastle team was intending to ride the wave of publicity for the South Korean paper, or if it simply submitted its paper to the journal at a fortuitous moment. And in an ideal world, science reporters would know the difference between a significant breakthrough and a local, incremental result.

But the premature release of this incomplete information, without any form of peer review and without making it clear to journalists that the work had not been refereed, is contrary to good scientific practice. The paper could, in principle, be revised or even rejected after peer review, in which case the public would have been misinformed. The absence of a paper also prevents other researchers from assessing or responding to the Newcastle results.

Industrial companies already release claims to the media while keeping data confidential for commercial reasons, and that's frustrating enough. The last thing the science community needs is for publicly funded academic researchers to start playing the same game. And to do so in the technically and ethically contentious arena of stem cells is playing with fire. ■

"The premature release of incomplete information, without peer review and without making it clear to journalists that it has not been refereed, is contrary to good scientific practice."

RESEARCH HIGHLIGHTS

Baby blues

Proc. R. Soc. Lond. B doi:10.1098/rspb.2005.3057 (2005); *Biol. Lett.* doi:10.1098/rsbl.2004.0274 (2005)
Prime examples of a biological trade-off have been discovered in two species of bird: caring for too many offspring in one year limits reproductive success the next.

On Norway's frozen north coast, a team led by Sveinn Hanssen of the University of Tromsø added eggs to the clutches of female common eiders (*Somateria mollissima*). The birds lost so much weight while sitting on the clutches, which were at the extreme of the usual sizes, that their fecundity dropped the following year.

Meanwhile, researchers from Auburn University in Alabama gave male eastern bluebirds (*Sialia sialis*, pictured) large broods of young to look after and found that they struggled to produce their trademark plumage the following season. Birds with a lighter parental load sported much brighter finery and enjoyed better mating success.

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

DAYBREAK IMAGERY/OSF

QUANTUM PHYSICS

Dancing vortices

Phys. Rev. Lett. **94**, 190401 (2005)

Pushing around vortices in a sea of ultracold atoms could shed light on how phase transitions work in superconductors, suggest theorists led by Nick Bigelow of the University of Rochester in New York.

Cooling and confining a cloud of gas atoms can force them to form a type of matter known as a Bose–Einstein condensate. Start this substance swirling and the vortices that form tend to settle into a regular static pattern. But Bigelow predicts the vortices can be pushed through a series of different arrangements by a grid of laser beams. The pattern changes as the beams brighten. The sudden rearrangements would be similar to structural phase transitions in other systems — including those that can cause power loss in superconducting wires.

NEUROSCIENCE

Patch work

Nature Neurosci. doi: 10.1038/nn1474

The difficulties dyslexic people have in reading may stem from a poor ability to detect visual signals through background noise, suggest Ann Sperling, now at Georgetown University in Washington DC, and her colleagues.

They showed 28 dyslexic and 27 non-dyslexic children two patches of visual noise

similar to television static; in one patch, the noise was superimposed on a particular pattern. The dyslexic children needed greater contrast to spot which patch hid the pattern.

The group's findings contradict the popular hypothesis that defects in the neurons of the magnocellular pathway — a connection between the retina and the brain — are responsible for deficits in visual processing in dyslexia. The magnocellular pathway is in the clear because the dyslexic children demonstrated the same need for greater contrast when shown patterns designed to stimulate a different visual pathway.

MALARIA

Caught in the act

PLoS Biol. **3**, 192 (2005)

The liver-infecting action of the malaria parasite has been caught on camera for the first time.

Ute Frevert from the New York University School of Medicine and her colleagues modified the parasite *Plasmodium berghei* so that it fluoresced. They then used a digital microscope to watch it invade the livers of live mice. Their recordings show the parasites moving into the livers' Kupffer cells, confirming suspicions that these phagocytic cells are their entry point. Kupffer cells normally digest and demolish anything foreign. Frevert's images prove that the malaria parasite must disable this response, allowing it to spread to other liver cells and multiply.

CELL BIOLOGY

The matrix: unloaded

J. Cell Biol. **169**, 1–11 (2005)

A family of enzymes once thought to encourage tumour development could also inhibit it. A team at the University of California, Los Angeles, has found that MMPs, or matrix metalloproteinases, restrict the growth of new blood vessels in a tumour.

Cancerous tumours grow their own blood supply by making a protein called VEGF, which encourages blood vessels to form. The VEGF sticks to the matrix surrounding cells, and the MMPs work by cutting it free. Surprisingly, the free VEGF does not promote the formation of blood vessels, but instead triggers the enlargement of existing vessels that do not support tumour growth.

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

BSIP, CHASSENET/SPL

MERCURY

Shell shock

Earth Planet. Sci. Lett. **234**, 27–38 (2005)

The Mariner 10 probe's encounters with Mercury in 1974 and 1975 showed that the planet has a magnetic field — a surprising find for scientists who expected the small planet's iron core to have completely solidified. Although studies have since shown that a thin liquid shell of iron may persist around the core, it was not clear whether such a shell could sustain the kind of circulation needed to generate a magnetic field.

Now researchers led by Sabine Stanley at Harvard University use computer models to show that convection in a shell can indeed produce magnetic fields similar to those observed at Mercury (pictured right). The work cannot rule out magnetic rocks in the planet's crust as the source of the field, but NASA's MESSENGER mission should be able to test the theory when it arrives at Mercury in 2008.

MEDICINE

Breathe easy

Science doi:10.1126/science.1108228 (2005)

Experiments in mice have revealed that nitric oxide metabolites play a role in preventing asthma, a condition suffered by more than 100 million people around the world.

A team headed by Jonathan Stamler of Duke University in Durham, North Carolina, engineered mice to lack the enzyme S-nitrosoglutathione reductase, leading to elevated levels of metabolites known as S-nitrosothiols. Even when exposed to allergens known to induce asthmatic symptoms in mice, the engineered animals were able to breathe easily, suggesting the metabolites help to keep their airways open. Medicines that inhibit the reductase enzyme might, therefore, be able to prevent asthma attacks in humans.

GENETICS

The bald truth

Am. J. Hum. Genet. (in the press)

Male-pattern baldness is a misfortune that men inherit mainly from their mothers. A systematic survey of the genomes of brothers from 95 families reveals that variability in the androgen receptor gene on the X chromosome is the single most important cause. The research, led by Markus Nöthen of the University of Bonn, Germany, adds to evidence from an earlier investigation of male-pattern baldness that implicated the same gene. As yet, the mechanism remains unclear.

IMAGE UNAVAILABLE FOR COPYRIGHT REASONS

RNA INTERFERENCE

Silent assassin

Nature Biotechnol. doi:10.1038/nbt1101 (2005)

Hopes are high for using RNA interference — a recently discovered gene-silencing tool — to treat disease. The challenge has been to get the RNA molecules into the right cells, and it seems that antibodies might work as delivery vehicles.

Judy Lieberman's group at Harvard Medical School fused an antibody that targets HIV-infected cells with protamine, a compound that binds nucleic acids. The researchers demonstrated that small interfering RNAs bound to this fusion protein are carried across infected cells' membranes, and retain their activity.

Lieberman's small interfering RNAs were designed to suppress a gene that HIV needs in order to replicate, and they even managed to inhibit the virus in cells known as T lymphocytes. Transferring foreign genetic material into these cells has been particularly difficult. In mice, the antibody complex selectively sought out cells expressing HIV proteins after it was simply injected into the bloodstream.

RADIOACTIVITY

Zinc's double decay

Preprint nucl-ex/0505016 at <http://arxiv.org> (2005)

The existence of radioactive isotopes that can decay through the simultaneous emission of two protons was predicted more than 40 years ago. But it was not until 2002 that the first such isotope, iron-45, was synthesized.

Now Bertram Blank and his colleagues at the French heavy-ion accelerator GANIL have produced a second unstable proton-rich element. They smashed a beam of nickel-58 into a nickel target and isolated eight nuclei of the isotope zinc-54 the resulting fragments. Seven of these nuclei decayed by means of the two-proton mechanism, with an estimated half-life of around 3 milliseconds, supporting the predictions of theoretical models.

ARTIST'S IMPRESSION/A. GRAGERA/SPL

JOURNAL CLUB

Philippe Janvier
National Museum of
Natural History, Paris

A specialist in early vertebrates retraces the discovery of an ancient fossil fish — his first impressions, later worries, and the relief prompted by a recent paper.

In 1999, my colleague Zhu Min showed me the first photographs of what he thought were two 535-million-year-old fish, found in Chengjiang, China. If they were actually fish, these fossils would be the earliest evidence of vertebrates — by about 50 million years.

My first reaction was that these small, leaf-shaped creatures did indeed resemble vertebrates. I wrote a News and Views article for *Nature* (**402**, 21–22; 1999) that explained how the fossils matched our conception of an ideal vertebrate ancestor. The fish seemed to combine aspects of the small, headless lancelet and the eel-like lamprey, which is one of the most primitive fish alive today. Could it be too good to be true?

Many similar specimens turned up soon after, providing information about the head and backbone of this early group, now known as Myllokunmingia. However, I became obsessively worried that all the recorded specimens lacked a tail tip. What if the tail looked like that of some of the strange, soft-shelled arthropods that also come from the Cambrian period? I began to fear that myllokunmingiids might not be fish — that we had been led astray by some extraordinary case of convergence.

So I was relieved when I read the first description of a complete myllokunmingiid tail in the *Journal of Evolutionary Biology* (X.-G. Zhang and X.-G. Hou **17**, 1162–1166; 2004). It is a good fish tail, pointed to the rear and bearing a vertical web that joins the tail to the dorsal fin.

Certain enigmatic features of myllokunmingiids remain, but I am happy that the question of their tail is settled.

NEWS

Flu in wild birds sparks fears of mutating virus

The deaths in China of more than 1,000 migratory birds from the flu strain H5N1 has left experts struggling to square the outbreak with their knowledge of the virus. At the same time, rumours are beginning to circulate that humans in the region have also fallen victim to the disease — although official sources have so far denied this.

The H5N1 strain has killed at least 53 people in Asia since late 2003, and is seen as one of the prime candidates for sparking a human pandemic. Migratory birds can act as carriers of flu, but their role in spreading highly dangerous strains such as H5N1 remains a matter for debate.

Until the latest outbreak, only a handful of migratory birds were known to have died from H5N1. This led some experts to suggest that the migrants are asymptomatic carriers of the virus, causing the occasional outbreak among poultry populations along their migration routes. Others believed that the small number of deaths among migrants were simply the result of wild birds picking up the infection from local ducks or chickens.

But the revelation on 21 May that at least 500 wild birds across five different species had died from the virus has dramatically altered the situation. With H5N1 now seeming to be highly infectious and lethal among the migrants, experts fear that the virus's genes may have mutated or reassorted.

To find out, the World Health Organization (WHO) is pressuring China to release samples for sequencing and analysis. "This is an

exceptional case," says Maria Cheng, the WHO's spokeswoman in Beijing. "We want to see the virus as soon as we can."

China had not reported any cases of H5N1 in people or birds since a previous poultry outbreak ended in June 2004. But several Internet sites including ProMED-mail, an online database of health-related news, are now reporting that people are dying as a result of the latest outbreak (see 'China rejects Internet claims of human cases', below). Some sources are claiming there have been up to 120 fatalities.

The Chinese health authorities deny that any human cases have been found. Since early May, when 178 suspicious waterfowl deaths were first reported in the Niannaisuoma village of northeastern Qinghai, authorities have quarantined the area, requiring people going there to wear goggles, masks and gowns. Increased surveillance so far has found no sick people, officials say.

There is no evidence of human to human transmission in the area, and large-scale transmission from so few wild birds directly to humans is unlikely, says virologist Vincent Deubel, who heads the Pasteur Institute in Shanghai. "A more virulent form for birds does not mean a more virulent form for humans," he explains. "It would be unrealistic to expect so many human deaths from an outbreak in only 500 birds."

But the numbers of reported bird deaths might hugely underestimate the actual toll, says Fusheng Guo, coordinator for the UN Food and Agriculture Organization's regional avian-

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

In a flap: a sudden outbreak of H5N1 flu in migratory birds has caught experts unaware.

V. YU/AP

flu surveillance network. "Many of them won't be found," he says, adding that his colleagues in Qinghai are reporting many more deaths every day. "I'm very worried about this." In fact, less than a week after the figure of 519 bird deaths was released, Jia Youling, director of the veterinary bureau at China's agriculture ministry, revised the number to more than 1,000.

Baoxu Huang, director of China's National

CHINA PHOTOS/GETTY IMAGES

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

Health workers disinfect an area in Qinghai.

China rejects Internet claims of human cases

Reports that the outbreak of avian flu among wild birds in Qinghai has resulted in any human cases of the disease are being denied by China.

Rumours that people had been infected began on 25 May, when the ProMED-mail Internet alert system relayed a translation of an article from the independent Chinese website Boxun News, which is based in the United States. ProMED warned that the information was unofficial and needed to be confirmed.

It was an article on Boxun News that first led ProMED to report the outbreak of severe acute respiratory syndrome (SARS) in 2003. Until then, China had been able to cover up the presence of

the disease. This time, the initial Boxun report claimed six people had died of avian flu in Qinghai in May. As *Nature* went to press, Boxun was reporting as many as 200 cases and 121 dead, and detailing the locations.

The reports were swiftly rebuked on 26 May by Xinhua, the Chinese official press agency, which denied that any human cases had occurred. But the statement added that "hospitals in Gangca County, where the avian flu cases were reported, have opened up a separate outpatient department for feverish patients".

A spokesman for the World Health Organization (WHO) says that the agency has

**PLASTIC PERIL**

Mouse study suggests bisphenol A could raise risk of breast cancer.

www.nature.com/news

GETTY

UK research councils claim success for open-access publishing plan

LONDON

Britain's main public funders for research seem to have achieved the impossible — they've come up with a policy that pleases both sides in the debate over open-access publishing. But appearances can be deceptive. Behind public praise for the statement, some publishers are voicing fears that small journals will go out of business, which could put scientific societies at risk.

Opponents of the current system of scientific publishing have lobbied hard in recent years, calling for all publicly funded research to be made available in free-to-access journals or archives. Their campaign, which in the United States included television adverts, has worried academic publishers. The fear is that libraries will cancel their subscriptions if papers are made available for free.

Supporters of open access are claiming victory in the wake of rules drawn up by Britain's research councils, which distribute most government science funding. The policy has delighted them because it requires all council-funded papers be put in an open-access archive "as soon as possible" after publication. Other major funders of research around the globe, including the US National Institutes of Health (NIH), allow researchers to wait up to a year before depositing their work.

Stevan Harnad, an advocate of open access and a cognitive scientist at the University of Southampton, believes that the UK policy's insistence on submission will make the use of open-access archives a regular part of academic life. "Once the history of this is written, this statement will be the single most important factor," he says.

But a crucial change to the policy, made following complaints from publishers, could dilute the power of archives. After consulting on an initial draft issued last autumn, the councils changed the policy so that submissions to archives will be subject to the copyright and licensing arrangement of the journal publishing the paper. Publishing executives say privately that they can now rewrite their rules so that submission takes place after a delay of several months, which will

The UK policy

Scientists will submit papers to subject-specific archives or to an equivalent run by their institution. The paper would only be the final text document accepted for publication, not the formatted version that is printed.

If this causes a range of archives to proliferate, access to papers should still be straightforward. Scholarly search engines, such as that unveiled last year by Google, automatically look through institutional repositories, so users shouldn't need to know where an article is actually held.

protect their subscription revenues.

Commercial journals are happy with the policy, but other publishers remain fearful. Learned societies, for example, often fund activities such as fellowship schemes through publishing. They say that libraries are strapped for cash and will consider cancelling subscriptions once archives take off, especially for journals that publish only a few times a year.

"We simply don't know how much damage this will do," says Sally Morris, chief executive of the Association of Learned and Professional Society Publishers, based in West Sussex. "If other funders follow this route, much material will be made available for free. Why would you pay?"

Harnad says there is no need to worry, as fields in which archiving is common, such as computer science and physics, show no evidence of failing journals.

The Wellcome Trust, Britain's biggest medical charity, is even more bullish about the idea. It said on 19 May that all papers produced using its money will have to be submitted to the NIH archive PubMed Central or to the British equivalent that is being developed. "Old journals sometimes cease to publish, but new ones spring up," says Mark Walport, the trust's director. "I have some sympathy with the learned societies, but it is not the primary mission of funders to support them."

The councils' statement still has to be "fine-tuned", say officials. Originally due for release this month, it has been put back until the summer, but is not expected to undergo significant changes before then. ■

Jim Giles

Animal Quarantine Institute, says that the Qinghai cases are probably isolated. He adds that China has more than 450 surveillance sites looking for the virus in various animals, including migratory birds. But one source says staff at the sites lack the necessary resources and training, and are likely to be missing cases. The surveillance stations in Qinghai, the source says, have machines for diagnosing samples, but people there don't know how to use them, and don't have the necessary reagents. ■

David Cyranoski

received written confirmation from the Chinese Ministry of Health. "The WHO does not have any reason to believe that the ministry's report is not true," he adds.

An editor at Boxun says he reported the human cases in the hope that it would force an investigation. "I feel that something serious has happened. It is worth people paying attention to what is going on there," he says. "Remember, SARS happened in big cities, and China kept it secret for a few months. In a remote area, it is even more difficult to get the truth out."

Declan Butler

SPECIAL REPORT

Back in the race

An almost unthinkable defeat for President Bush in Congress has put embryonic stem-cell research firmly back on the US agenda. But with South Korea setting a pace the United States will still struggle to match, the field's future is fraught. *Nature* reports on the key political battles surrounding this issue.

Advocates of stem-cell research in the United States have just secured a victory that few believed was possible. On 24 May, the House of Representatives passed a bill that would allow federal funding for research on newly derived human embryonic stem-cell lines. If signed into law, it would reverse the policy set by President George W. Bush on 9 August 2001 that prohibits federally funded research on embryonic cell lines derived after that date.

Sponsors of the Stem Cell Research Enhancement Act convinced 50 members of Bush's Republican party to vote against him, in favour of the bill. A few months ago that seemed unthinkable, but a combination of medical and economic arguments has helped to bring about a change of heart.

Many representatives say that they were swayed by constituents or relatives with diseases that could potentially be cured by stem-cell research. Others gave credence to arguments

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

Unmoved: George Bush has said he will veto fresh stem-cell legislation.

that, by not funding the research, the United States is losing ground to other countries, such as South Korea. Just days before the vote, Woo Suk Hwang at Seoul National University announced a major breakthrough in stem-cell research — the use of cloned embryos to produce stem cells that genetically matched patients' tissues (see *Nature* **435**, 393; 2005).

Some representatives also expressed concerns that states such as California and New Jersey, which are contributing their own money to fund stem-cell research, may gain prestige and economic advantage at the expense of other states. Indeed, the vote is expected to bolster California's stem-cell initiative, which has become bogged down lately in political battles of its own (see 'California stem-cell institute fights legal challenges', below).

And the argument that many embryos left over from *in vitro* fertilization are discarded every year also seems to have been convincing. Under the new measure, scientists would be able to derive stem-cell lines from these leftover embryos, if the couple involved consented.

"The surplus embryos are going to be discarded as medical waste, and the notion that these may have the potential to cure dreaded diseases just resonates with people," says Pat White, director of federal relations at

California stem-cell institute fights legal challenges

Stem-cell research may have won a victory in the US Congress, but California's nascent \$3-billion programme is embroiled in a political battle of its own.

A coalition of state legislators wants to increase public control over how the California Institute for Regenerative Medicine (CIRM) will spend its money.

But last week, CIRM leaders hit out, saying that the proposals would make it "extremely difficult if not impossible" for the agency to function.

The CIRM, created by a statewide vote last November, is California's answer to the Bush administration's rules limiting the number of stem-cell lines that can be used for federally funded research. By selling state bonds, it

will provide about \$300 million a year over the next decade to drive stem-cell research that is not eligible for federal money.

The CIRM's leaders are eager to get the money flowing. "California is already behind," says Zach Hall, the institute's interim president. "We need to catch up and move ahead."

But there are concerns about the level of transparency and accountability in the CIRM's allocation of grants. Democrat Deborah Ortiz, a long-time advocate of stem-cell research and state senator from Sacramento, has introduced a bill to address this (see *Nature* **434**, 427; 2005). She and her coalition argue that without tight controls, critics of stem-cell research could attack the programme in future. Her original

bill, introduced in March, would have made public the peer-review process for grants, introduced stringent conflict-of-interest rules, and required repayment of all state costs, including grants, from the royalties of any drugs or products arising from CIRM-funded projects.

The CIRM's leaders say that this goes too far, and on 23 May the institute's board voted unanimously to reject the bill. "The law would stop us in our tracks," says Hall.

Later in the week, the board released a letter signed by the presidents of the state's main universities that argued the bill would discourage industry from working with the CIRM.

Ortiz has since removed some of the more contested provisions. But as the bill hit the California Senate

floor this week, it still required peer reviewers to release summaries on grant decisions. And the issue of paying back costs is also unresolved.

The Senate is expected to pass the bill, which would then go to the state assembly. If it passes there before 30 June, Californians will vote on the law either in November or in June 2006.

The CIRM is also facing a lawsuit filed against the state by a group that fights taxes. California cannot sell bonds to fund the institute's research until this suit is resolved. That means no one knows when the first grants will be available, although the CIRM hopes to give out \$15.3 million in training grants this autumn. These will be funded by donations until the bonds can be sold.

Rex Dalton, San Diego

C. DEVORAH/WENN/NEWS.COM

the Association of American Universities.

Advocates are now looking ahead to the fight in the Senate. The day after the stem-cell bill passed the House, six Republican and Democrat senators claimed that they have enough votes to get the bill through. If the bill is passed there, it still needs to be signed into law by President Bush. He has said twice that he will veto it, although that would be an unprecedented step — he hasn't yet used his right to veto a bill.

But even if he changes his mind, or Congress manages to override a veto, the bill would leave the United States in a conservative position compared with countries such as Britain and South Korea. These nations allow stem-cells to

be derived from cloned human embryos, a technique often called 'therapeutic cloning'.

"It's really a very small step forward, particularly in contrast to what's going on in other countries and even other states," says John Gearhart of the Johns Hopkins School of Medicine in Baltimore, Maryland, who was part of a team lobbying members of Congress to pass the bill on the day of the vote.

But advocates are savouring the victory as a sign that public opinion is slowly shifting towards such research (see page 537). "The sense I get is that there is stronger and stronger support for this work," says Gearhart. ■

Erika Check, Washington DC

From Nobel ambitions to hunger strikes...

SOUTH KOREA

Woo Suk Hwang of Seoul National University, whose team last month reported that it had created 11 human embryonic stem-cell lines genetically matched to individual patients, has become a national hero. Politicians are pledging their support to help Hwang win a Nobel prize. And the government seems keen to capitalize on his research. Hwang's research budget from the science ministry jumped from 6.5 billion won (US\$6.5 million) last year to 26.5 billion won this year and is expected to rise further.

In an interview with *The Korea Times*, Ky Young Park, presidential adviser for science and technology, and a former co-author of Hwang's, says Korea is "mulling over a global consortium to study the next-stage technologies of differentiating stem cells into specific cells or organs". Hwang himself reportedly has plans to create a stem-cell 'dream team', gathering top names from labs in the United States, Britain and elsewhere to collaborate on turning the technology into cures.

ITALY

Thirty scientists have gone on hunger strike in protest at what they say is a distortion of scientific facts in a feverish referendum campaign. Research using human embryonic stem cells is completely banned in Italy, but the country faces a public referendum this month on whether to allow 'therapeutic cloning'.

Media coverage of stem-cell science is confusing the public, in particular by reporting that adult stem cells have the same medical potential as embryonic stem cells, the scientists say. They also complain that lobbies opposing human embryonic stem-cell research get disproportionate airtime. "Citizens have the right to decide according to their own ethics — but they have to be correctly informed about the science," says hunger striker Gilberto Corbellini, a medical historian and bioethicist at La Sapienza University in Rome.

GERMANY

Human embryonic stem-cell research is restricted to the use of cell lines created before 2002, but the topic has suddenly become an unexpected election issue. In response to Hwang's announcement last month, Chancellor Gerhard Schröder said that his government would be prepared to reconsider the law. But two days later his Social Democrat party lost a decisive regional election, and he called a surprise general election for September.

Parties are now lining up on either side of the ethical divide. Schröder's research minister Edelgard Bulmahn has come out clearly in favour of a parliamentary debate to reconsider the current restrictions. The major opposition CDU party is against change. But the CDU's likely coalition partner in any future government, the FDP, has spoken strongly in favour of therapeutic cloning.

FRANCE

France's bioethics law forbids therapeutic cloning, but revisions drafted by Roger-Gérard Schwartzberg, the research minister in the former socialist government, allow research on any human embryonic stem-cell line for a five-year period. He now wants to go further. On 24 May, he introduced a draft law that would allow therapeutic cloning, saying that the Hwang paper "marks a decisive step towards regenerative medicine".

AUSTRALIA

The country currently allows embryos left over from *in vitro* fertilization to be used as a source of stem cells, but bans both reproductive and therapeutic cloning. The law is under review, and advocates of stem-cell work are hoping the momentum created by the Korean discovery could tip the balance towards allowing therapeutic cloning.

Alison Abbott, Munich
David Cyranoski, Tokyo

Protein structures hint at the shape of things to come

In less than a year, an Anglo-Canadian group has worked out the structures of 50 complex proteins that are relevant to human disease.

The recent explosion of genome sequence data led scientists to hope that working out protein structures could become automated and accelerated. To date, projects attempting this have faced criticism for taking too long to deliver. But the results from the Structural Genomics Consortium (SGC), which specializes in human proteins, mark the coming of age of the field. The group, which is based in Toronto and Oxford, says it will add a further 100 structures to its free-access database in the coming year.

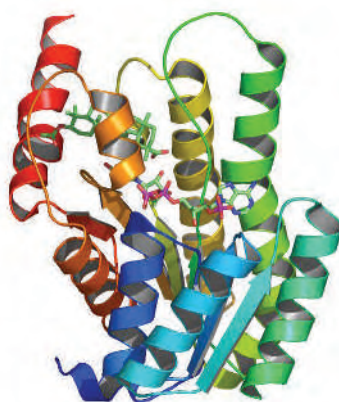
Knowing the shape of a protein is crucial for understanding its biological function, and designing drugs to interact with it. For example, the SGC has just released the three-dimensional structure of a human enzyme (pictured) that converts cortisone to the metabolic hormone cortisol.

"Mice that lack the gene for this enzyme do not develop diabetes, whatever their diet," says Aled Edwards, the consortium's executive director. "So this target is exciting the pharmaceutical industry at the moment."

The field of structural genomics aims to reveal the shapes of proteins by using the information obtained from genome-sequencing projects in automated, high-throughput schemes. These would synthesize thousands of proteins *in vitro* so that they could be examined by X-ray crystallography and nuclear magnetic resonance techniques, which are used to image molecules in three dimensions.

Techniques have come a long way since Max Perutz took 22 years to determine the first structure of a protein, haemoglobin (M. F. Perutz *et al.* *Nature* **185**, 416–422; 1960). And increasing availability of genome sequence information in the past few years caused structural genomics initiatives to pop up all around the world.

Apart from the Anglo-Canadian group, Germany, Japan and the United States all started major structural genomics projects. The field got off to a slow start, and the ambitious, Berlin-based Protein Structure Factory, which was investigating human proteins, closed after it failed to secure more funding in 2004. But the rate at which structures are



SGC

Mortal coils: knowing the shape of enzymes is vital — this one could guide research in diabetes.

being published is starting to accelerate.

The US National Institute of General Medical Sciences, based in Bethesda, Maryland, started its Protein Structure Initiative (PSI) in 2000. By this summer, the PSI's nine 5-year projects will have delivered 1,100 structures — nearly half of them produced in the past year.

And in a few weeks' time, the institute will announce the names of labs that have been selected to participate in the PSI's second phase. This will aim to deposit hundreds of protein structures in public databases each year.

But despite the impressive number of structures produced, the PSI has focused mainly on bacterial proteins. These are simpler than mammalian proteins and are therefore easier to make and purify. The SGC project is the first to solve the structures of a large number of human proteins.

The consortium's achievement is saluted by the PSI's coordinator, John Norvell. "Every step in the process is harder for human proteins," says Norvell. "Cracking 50 structures in less than a year is very impressive."

Norvell says that the US National Institutes of Health is planning pilot projects that will work out faster ways of capturing similarly difficult proteins.

He adds that after several years of building technologies, structural genomics initiatives are ready to roll out protein structures in large numbers. ■

Alison Abbott

"Cracking fifty structures in less than a year is very impressive."

ON THE RECORD

"You have to admit that evolution theory is not complete."

Dutch education minister, Maria Van der Hoeven, tells newspaper *De Volkskrant* why she supports a public debate on intelligent design.

"I am very, very tired of the US signing on to international science agreements that we later come to regret."

Congressman Sherwood Boehlert (Republican, New York) supports an amendment to stall US involvement in the multibillion-dollar fusion experiment ITER until a clear source of funding can be found.

"We need to ban the sale of long pointed kitchen knives."

The *British Medical Journal* on how to reduce knife crime.

THREE GOOD REASONS...

...to make your next car a hydrogen car

Mileage Last week a prototype hydrogen car achieved a whopping 3,836 kilometres per litre of fuel.

Convenience for Californians. The state has just pledged to open 100 hydrogen-fuel stations around Los Angeles and San Francisco ... by 2010.

Looks General Motors is developing a stylish hydrogen-powered Hummer.

NUMBER CRUNCH

50 — the number of years since the first atomic clock was unveiled at the National Physical Laboratory in Middlesex, UK.

0.16 — the number of seconds by which that original clock would be off if it was still running today.

0.1 — the number of microseconds a modern atomic clock loses each year.

32 — the number of 'leap seconds' that have been added to atomic time since 1972 to adjust for a gradual slowing of Earth's rotation.

Yeast feeds debate on prolonging life

'Eating less helps you live longer' has for decades been the message from researchers of ageing. So experts are not sure what to make of a study in flies that suggests it could be what you eat, not how much you eat, that really counts.

The idea that restricting calories prolongs lifespan was first reported in 1935, following studies in rats. The observation has since been supported by studies in species ranging from worms to dogs. The source of the calories is generally considered irrelevant.

Now a research team from University College London has extended the lifespan of *Drosophila* flies by reducing the amount of yeast (a source of protein and fat) or sugar in their diets. The team, led by Linda Partridge, observed a much more dramatic effect with yeast than with sugar, even though the overall change in calorie content was the same. The results are published online in *PLoS Biology* this week (W. Mair *et al.* 3, 223; 2005).

The finding hints that reducing protein and fat might be the key to living longer, rather than cutting down on the total number of calories. But many researchers are sceptical of drawing any broad conclusions. "We already saw what a disaster it was in the 1990s with fad diets that lowered fats and increased carbs," says Leonard Guarente, a molecular biologist at the Massachusetts Institute of Technology in Cambridge. "That's when people really got fat."



Roll on the years: reducing calorie intake may help you live longer, but does it matter what you eat?

He thinks the results may simply reflect a peculiarity of flies. "You need to know how well flies metabolize glucose compared with yeast," he says.

Signe Zou, a fly geneticist at the National Institute on Aging in Baltimore, adds that other components in the yeast might also be having an effect. "Yeast are made up of a lot of

ingredients," he says. He would like to see the experiment repeated using pure protein or fat extracts.

Partridge's study shows how tricky it can be to pin down such effects, agrees Richard Weindruch, a gerontologist at the University of Wisconsin at Madison. "This solidifies my concerns about the nuances and difficulties of conducting studies of caloric restriction in some model organisms."

But researchers are sure, at least, that restricting calories does prolong lifespan, even if they do not know how. "The name of the game is not to take in more energy than you need," says Guarente.

How many calories would humans need to cut to gain years, or even decades? "Based on our mouse data, I'd guess a minimum of 20% from a predetermined baseline, for a person who is not obese," says Weindruch.

Weindruch is currently evaluating diet-restricted rhesus monkeys, in a study that started more than a decade ago. "The monkeys are now middle-aged and it is clear the diet is doing them good," says Weindruch. They are protected from type II diabetes, a common ailment in ageing monkeys, and seem to have healthier hearts. "I look forward to reporting the outcome," he says, "assuming the monkeys don't outlive me!"

Carina Dennis

Drug giants fail to name compounds in trial database

WASHINGTON DC

An international group of medical editors is challenging several leading pharmaceutical companies, saying that their reporting of clinical trials is deliberately incomplete.

The International Committee of Medical Journal Editors made their complaint in an editorial in *The New England Journal of Medicine*, published online on 23 May. They argue that leading pharmaceutical companies are obeying the letter but not the spirit of a 1997 law that requires the public registration of ongoing trials involving serious or life-threatening illnesses.

The government-maintained registry, www.clinicaltrials.gov, is intended to help patients find information about clinical trials. But the editors say that drug firms are inserting a "meaningless phrase" instead of the names of drugs, so patients aren't getting

the full picture, including any negative data.

The New England journal's editor-in-chief, Jeffrey Drazen, says that Merck, GlaxoSmithKline (GSK) and Pfizer in particular "didn't meet the sniff test" in a review conducted early this month by Deborah Zarin, the database's director. Zarin found that specific drug names were missing in scores of trials, which used the phrase "investigational drug" to describe their products. Drugs weren't named in 36% of 75 Pfizer studies reviewed, in 53% of 55 GSK trials, and in 90% of 132 Merck trials.

Drazen argues that patients deserve more. "It's not right," he says.

The drug companies insist that they are trying to make the reporting of results as transparent as possible. They claim they are complying with the law, which does not explicitly require companies to name drugs,

but asks them to describe the "intervention" being used. "We think we've made big strides in improving the transparency of clinical data. And we will continue to do so," says GSK spokesman Rick Koenig.

Pfizer's spokeswoman Betsy Raymond says her company withholds the names of certain drugs for competitive reasons. Merck did not return a call seeking comment.

The editors' committee wants to see trials being publicly registered in a meaningful way, partly so that negative results about particular drugs can be accessed. They have defined a list of minimum criteria that companies must provide. And this summer, the editors will start refusing to publish trials that do not register this information. The editorial "is a message that we are paying close attention," says Drazen.

Meredith Wadman

**ANOTHER LEVEL**

Researchers home in on the part of the brain that unravels the meaning of metaphors.

www.nature.com/news

CENTER FOR BRAIN & COGNITION

US treasury seeks bright ideas to beat bogus dollars

WASHINGTON DC

Threatened by cheap, high-quality laser printers and other technologies producing counterfeit currency, the US Department of the Treasury has asked leading scientists to come up with ways to defeat phoney bills.

The National Academy of Sciences applied itself to the subject in 1993, when it urged the treasury to use features such as watermarks and colour-changing ink. These were duly incorporated in bills, but are now being faked in an increasing number of counterfeits.

Last year, \$43 million in counterfeit money was removed from circulation in the United States. "Counterfeiters are getting awfully good, especially with all the digital technology," says Goutam Gupta, head of technical support at the treasury's Bureau of Engraving and Printing.

Now an academy panel is pondering how to fight back. The newly appointed group includes experts in biomaterials, optics and



Face lift: dollars may soon look quite different.

neuroscience, and may explore avenues such as nanotechnology. Any new element on bills would have to be difficult to simulate, but cheap to mass-produce and easy to recognize.

For example, the euro quickly became a target for counterfeiters, because it is so widely used. But with features such as holograms, iridescence, watermarks and a security thread, it is faked less than the currencies it replaced.

"We want something that will attract a person's eye," says Lenora Clarke of the bureau, which pumps out almost \$700 million each day.

Not all cashiers look carefully enough at bills to notice something as subtle as the absence of a security strip, says Lisa DiNunzio, programme manager within the bureau's office of security technology. The most useful feature, according to bank tellers, is the feel of the bill. Unlike more usual wood-based paper, dollars are made of cotton and linen.

But counterfeiters can get around this by bleaching off the markings on another kind or denomination of currency. The US Secret Service has found \$100 bills printed on bleached-out Iraqi dinars, for example, with the original watermark of a horse faintly visible over Ben Franklin's shoulder.

The panel expects to finish its work within two years, but hopes to issue interim recommendations by February 2006.

Emma Marris

K. SIMPSON/NATURE

D. MCNEW/GETTY IMAGES

IMAGE UNAVAILABLE FOR COPYRIGHT REASONS

Ready for action? It is unclear whether the WHO will investigate suspected bioterror attacks.

WHO sets health rules but ducks bioterror issue

New guidelines governing the conduct of the World Health Organization (WHO) have been agreed — but it's still not clear whether the agency should have a role in handling suspected bioterror attacks.

In Geneva on 23 May, the 192 WHO member states accepted a final version of a document called the International Health Regulations, which govern how the agency manages disease outbreaks. Years of negotiations over the document had stalled because of arguments over the WHO's powers to investigate bioterrorism.

Some negotiators had pushed for the agency to take on the role of policing possible violations of international biological-weapons treaties. But several nations were uneasy about providing the WHO with the sensitive information needed for it to act effectively.

The final version of the document eliminates any mention of bioterrorism, but requires states to cooperate with the WHO in "public health emergencies of

international concern". Observers say the WHO has gained some power through this clause, but that there is still much uncertainty about how the international community will deal with bioterrorism outbreaks in the future.

Union scraps boycott of Israeli universities

A British academic teachers' union has abandoned plans to boycott two Israeli universities.

The original decision by the Association of University Teachers to sever links with the universities of Bar-Ilan and Haifa was made in April as a protest against Israeli policies in the Palestinian territories, but prompted an angry reaction from academics. A special meeting of the union's council, held on 26 May in response to the outcry, agreed to scrap the boycott. Union officials say they will focus instead on providing support to Israeli and Palestinian academics.

But the reversal angered Palestinian academics, some of whom backed the boycott as a means of pressuring Israel to end its occupation of the West Bank and Gaza (see *Nature* 434, 813; 2005).

Ireland finds transgenic corn after US tip-off

The Irish authorities last week impounded a US shipment of animal feed containing a strain of genetically modified corn that is banned in the European Union. This provides early evidence that a new testing regime seems to be working.

It is the first such catch from some 290 shipments examined so far. US inspectors test any corn destined for European markets,

and alerted their Irish counterparts when the test came back positive. The corn was intercepted at the port.

The European Commission insisted on the test system after *Nature* revealed in March that the Swiss biotechnology firm Syngenta had inadvertently released hundreds of tonnes of unauthorized seeds of a transgenic form of corn, called *Bt10*, onto the market. The news caused particular concern because *Bt10* contains a gene that confers resistance to the antibiotic ampicillin (see *Nature* 434, 548; 2005).

The Irish authorities are holding the corn pending a decision on its disposal.

Nuclear treaty meeting leaves delegates up in arms

A conference to determine the course of the world's leading nuclear arms-control treaty has ended in stalemate.

Delegates from more than 180 nations gathered in New York last month to discuss possible adjustments to the Nuclear Non-Proliferation Treaty, the main global accord to stop the spread of nuclear weapons. The meeting was the seventh such review since the treaty came into force in 1970.

Arms-control experts had hoped that the delegates would use the meeting to strengthen the power of international inspectors, and that countries with nuclear weapons would recommit to reducing their arsenals (see *Nature* 435, 132; 2005). But after four weeks of acrimonious debate, attendees have little to show for their effort, says Rebecca Johnson, director of the London-based Acronym Institute for Disarmament Diplomacy.

Iran held up efforts to reach a statement on the Middle East, and the United States objected to language strengthening the disarmament sections of the treaty.

Freedom of choice leads Italy to failed cancer drug

Italy's government is flirting with paying for a controversial cancer therapy that clinical trials have shown does not work.

The treatment, a cocktail of natural products including the expensive drug somatostatin, was developed by physician Luigi Di Bella. In 1998 the government gave in to public pressure and agreed to sponsor clinical trials, but these provided no evidence of beneficial effects, and interest in the therapy subsided.

But last month, Francesco Storace, newly appointed health minister in Silvio Berlusconi's right-wing government, announced plans to consider fresh trials, saying that "freedom of choice for patients must be guaranteed".

K. KAMASHIDA/GETTY IMAGES

Nations line up to sink Japan's whale catch plan

Japan's plan to increase the number of whales it can catch for 'scientific' research purposes is already rousing some strong opposition from governments and non-governmental organizations.

At the International Whaling Commission's annual meeting this month, Japan will ask whether it can double its catch. Joji Morishita, a senior member of the fisheries agency's international-affairs division, says there are plenty of

IMAGE UNAVAILABLE FOR COPYRIGHT REASONS

scientific reasons for the increase. But he adds that the commission's rules prohibit him from discussing details before the plenary session starts in Ulsan, South Korea, on 20 June.

This hasn't stopped critics from responding to early

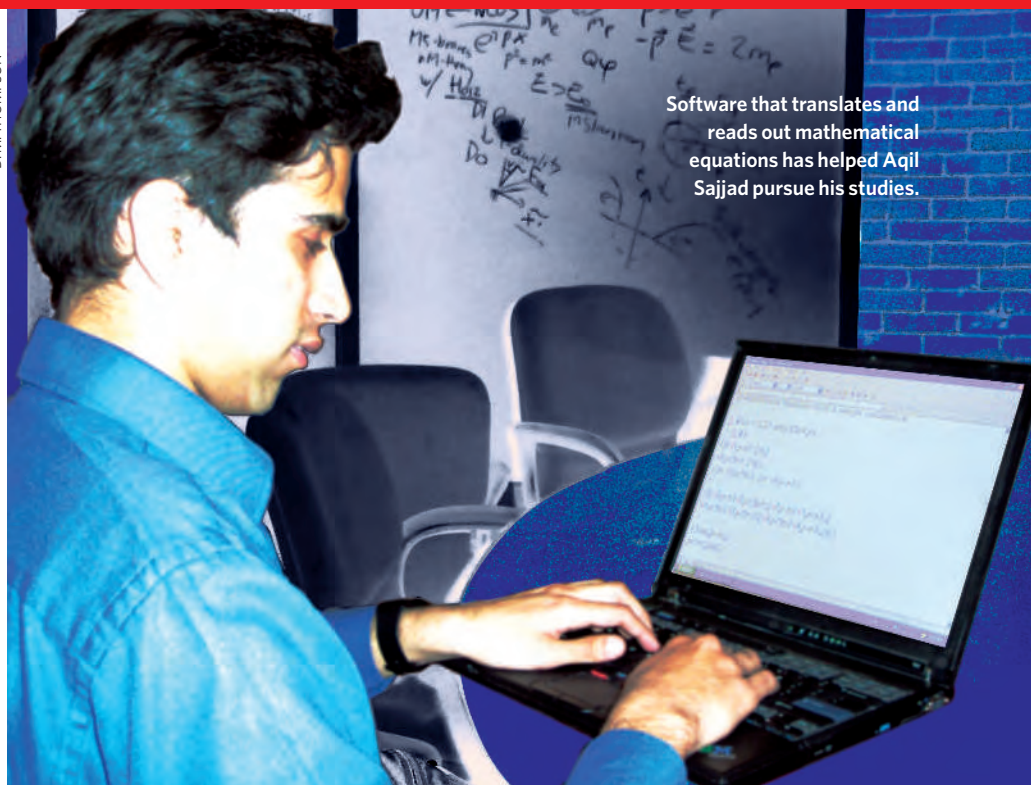
reports that Japan plans to catch humpback and fin whales, in addition to the four other species currently taken. Australian Prime Minister John Howard has written directly to his Japanese

counterpart, Jun'ichiro Koizumi, asking Japan to drop any such plans.

Morishita says that such pleas are based on emotion rather than science. "We want a scientific and constructive discussion," he says.

Access all areas

D. M. THOMPSON



Software that translates and reads out mathematical equations has helped Aqil Sajjad pursue his studies.

Scientific research can be tricky at the best of times, but people with disabilities face additional challenges both in the lab and when dealing with data. **Jessica Ebert** meets the researchers who are building their own customized solutions to overcome these problems.

"Russia...Sweden...Denmark." A mechanized voice calls out the names of European nations as Stan Berman moves his fingers delicately over the raised dots on a piece of thick white paper. The paper lies on a touch-sensitive pad attached to a laptop computer. A map of Europe, formed in outline by small, raised circles, is faintly visible on the paper. Berman, a blind business and technology consultant who provides services for the visually impaired, sits quietly with his eyes closed and head tilted slightly back as the words continue to stream from the computer's speakers. "United Kingdom...Ireland..."

"This is quite neat," he says. Berman and his colleague, Dan Grauman from the National Cancer Institute in Bethesda, Maryland, are sitting in a sunlit kitchen in Columbia, Maryland, to test-drive a gadget known as IVEO — one of the newest assistive technologies for the blind.

IVEO is the brainchild of John Gardner, a solid-state physicist turned entrepreneur who is also blind. It is the latest in a long line of products that he has invented to remove barriers that prevent the visually impaired from fully appreciating maths and science. Like many researchers with disabilities, Gardner had to develop his own technologies for doing science and communicating his work because commercial solutions were seldom available.

"I have a very strong philosophy that we all ought to be reading the same things," says Gardner. "When there isn't a way to do it, we make a way."

Virginia Stern, director of the project on science, technology and disability at the American Association for the Advancement of Science (AAAS) in Washington DC, attributes the advances in assistive technology to the can-do attitude of people with disabilities such as Gardner. "Assistive technology has dramatically changed opportunities for students with disabilities," she says. "It's been the push of the users, some of whom have developed things because they knew what they needed."

Gardner grew up blind in the left eye. Although he also had poor vision in his right eye, "It was correctable enough so I could drive — or as my friends say, 'aim' — a car," he says.

A difficult lesson

Gardner's early physics career took him from Texas to Illinois and then to Munich, Germany, where he studied the nuclear magnetic resonance of liquid copper alloys. In 1973, after six years on the physics faculty at the University of Pennsylvania in Philadelphia, he headed west to Oregon State University in Corvallis.

In 1988, Gardner had an operation to treat the glaucoma in his right eye. But rather than slow the loss of vision in that eye, complications left him prematurely blind. "It was not a pleasant time," he recalls.

Although he was away from the lab for a few months, he stopped working for only a couple of days. "I had so many graduate students and postdocs and grants and proposals in the works, I wasn't allowed to be sick," he smiles.

When he finally returned to the lab, he faced a new set of challenges. "I didn't know how to be blind," he explains. "No one knew what to do with me."

Initially, Gardner's students would tape-record journal papers for him and try to describe what their data looked like by taking his finger and tracing the curves in a graph or a picture. It took him months to discover that blind people could use computers. "I had a computer that was just sitting and gathering dust in my office," he says.

He can still recall the time he was introduced to screen readers, the software that converts what is displayed on a computer screen into either audio or a Braille output that can be read on a special keyboard. "That was one of the better days," he says.

From that moment on, Gardner began inventing tools that allowed him to continue his lab work. He never planned to abandon physics in pursuit of his innovations, but there was such an unmet need for the tools that he found himself moving in that direction. "I didn't really want to become an entrepreneur," he says, "but someone had to."

In the early 1990s, Gardner was awarded a grant from the National Science Foundation (NSF) to establish the Science Access Project at Oregon. In the past ten years, the project has developed technologies to overcome the biggest barriers for those with poor vision: reading and writing mathematical and scientific notation, and viewing graphs, tables, charts and diagrams.

Breaking down the barriers

"Helen Keller once said that blindness isolates you from things, but deafness isolates you from people," says Jane Dillehay, a deaf molecular biologist at Gallaudet University in Washington DC, which caters exclusively for deaf people. "Science is a communal enterprise and depends on people sharing their research with each other." The largest barrier for a deaf scientist is engaging in the spontaneous conversations that occur around the lunch table or in the hallways at conferences.

At conferences, the assistance of sign-language interpreters is really important, says Caroline Solomon, a marine biologist at Gallaudet. If

interpreters are not available, she uses remote captioning services to follow lectures. This is fine in the lecture hall, but Solomon dreams of a portable device to facilitate two-way conversations between a hearing and a deaf person. People are working on such tools, but "the technology cannot handle the average person with an unpredictable vocabulary and it is also not yet portable", says Dillehay.

The Internet has had the biggest impact on dialogue between the deaf and hearing. "I think the popularization of e-mail has been the greatest accessibility tool for the deaf community," says Derek



Jane Dillehay uses sign language to teach biology students.

Braun, a professor of molecular biology at Gallaudet who also conducts research at the US

National Cancer Institute. "In fact, some of my collaborators probably do not realize that I am deaf." **J.E.**

Today, Gardner is demonstrating IVEO to Grauman and Berman. Grauman's job as an information technologist at the National Cancer Institute includes making sure that maps showing cancer mortality across the United States can be used by everyone. Since 1998, federal agencies have been required to make their information and technology, including websites, operating systems and kiosks, accessible to people with disabilities.

All mapped out

Although the maps Grauman has developed meet federal requirements, problems remain. "They don't give the blind user the relative location of the states," he explains. IVEO is expected to change that. "The beauty of the touchpad is that it will permit the blind user to move east and west, or north and south, and actually 'see' whether there are neighbouring states with high cancer mortality rates," he says.

Universal access to the products of scientific research — from public-health information to data on environmental

pollutants — is just one aspect of assistive technology. But for students with disabilities, having access to the right technology can determine whether they choose to enter science at all. That was true for Aqil Sajjad, a physics student from Pakistan, who says that the specialized software WinTriangle, which helps the visually impaired read and write mathematics, was a lifeline.

Sajjad lost his eyesight in 1996 at the age of 16. Although science had intrigued him from a young age, he couldn't get the support he needed to study physics in his homeland. He began searching the Internet for educational opportunities in the United States because he knew the 1990 Americans with Disabilities Act had opened new doors for students with disabilities.

At some point during his search, Sajjad stumbled on the work of Gardner's Science Access Project and knew that this was where he wanted to pursue his passion for physics.

"It was crucial for my decision that Gardner

was a blind physicist and was developing these tools," he recalls. "WinTriangle turned out to be the thing I needed."

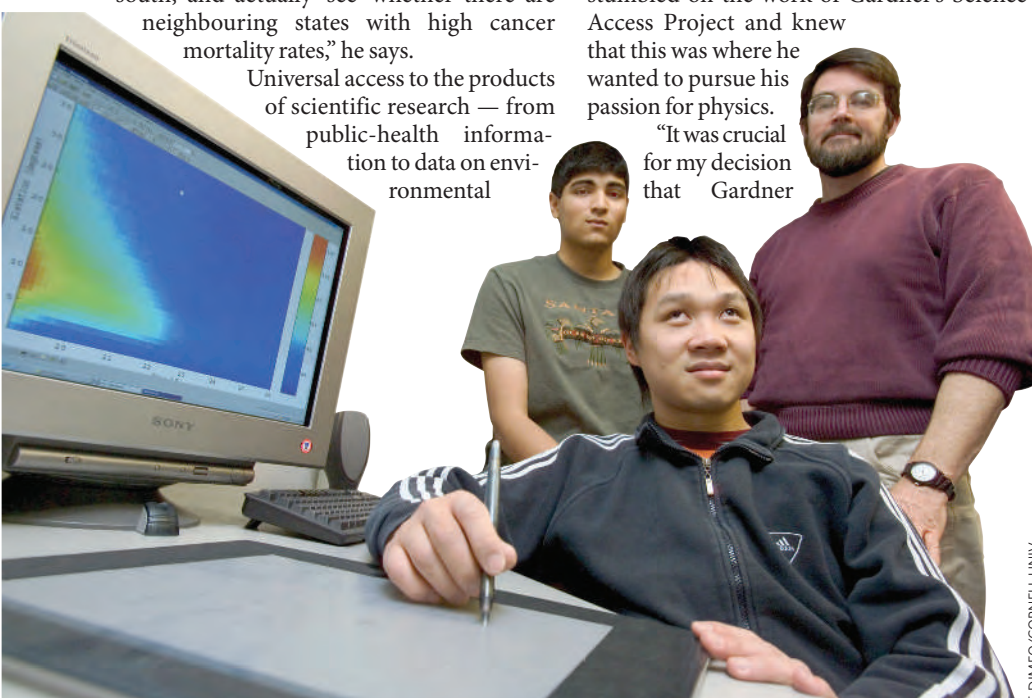
WinTriangle, which was produced by the Science Access Project, allows a blind user to write mathematical equations, perform calculations and hear mathematical text. In its simplest form it is a scientific word processor with a specialized set of fonts representing symbols and operations that can be read by a speech synthesizer.

Sajjad has now moved on and is a first-year doctoral student in theoretical physics at Harvard University in Boston, Massachusetts. Here, physicists write papers using a typesetting language called LaTeX, so David Thompson, a fellow student, wrote a software program to convert LaTeX into WinTriangle. "The concept behind this is so simple," says Thompson, but the significance for Sajjad is enormous. "We're completing the path of communication of maths between blind and sighted communities," he explains.

Open access

Thompson and the Adaptive Technology Laboratory at Harvard are now ironing out some of the glitches in the program and adding features to the converter to make it more robust, such as improving the quality of the voice synthesizer and adding more keyboard shortcuts. "Having the converter makes a huge amount of information available to me," says Sajjad. "It has made a big difference in terms of levelling the playing field between me and other people."

WinTriangle is 'open-source' software, which lets users adapt and rewrite it to meet their needs. It is fairly common for assistive technologies to be modified or enhanced by their users. Gardner recently teamed up with



F. DIMEO/CORNELL UNIV.

Tuned in: the colours on the map of the upper atmosphere can be turned into musical notes for blind student Victor Wong (front), thanks to work by Ankur Moitra (rear left) and James Ferwerda.

H. WATERS

Victor Wong, James Ferwerda and Ankur Moitra at Cornell University in Ithaca, New York, who have developed software that translates colour pixels on a computer screen into piano notes. The group hopes to combine the audio software with IVEO's tactile technology to solve a particularly challenging information problem.

Wong, who lost his sight in an accident when he was seven, works with a team that studies the ionosphere, the layer of the atmosphere between Earth and space. Part of Wong's work involves reading maps of the ionosphere in which colours represent variables such as electron density and light intensity. "To be able to visualize a map is a very basic need," says Wong. "There is no way to handle the image problem right now."

With the Cornell system, Wong uses a touch-sensitive tablet to explore three-dimensional images with a 'wireless' electronic pen. "When you move the pen around on the tablet it's the same as looking around on the screen," explains Wong.

When using the tablet, Wong finds the pen is almost too sensitive — it is actually better than the naked eye and gives too much detail. In addition, there is no easy way to uncover the major features of an image without tedious, pixel-by-pixel exploration.

Sound investment

This is where Gardner comes in. By combining IVEO's ability to display tactile maps with the Cornell team's software for converting colour into sound, users will be able to determine the boundaries of a map or graph more easily. "This is about finding a way to make graphical information accessible to the blind and those with print disabilities," says Gardner. "That is still one of the biggest barriers."

But not all of the barriers are technological. "Widespread use of these technologies requires both money and cultural adjustments," he says. "The latter is the hardest to achieve."

Stern agrees. "Counsellors and teachers at all levels, from preschool through to the very critical high-school years, do not believe that students with disabilities can persist and excel in science and engineering fields," she says. "One reason is there aren't enough role models." In addition, she says, few students, counsellors, teachers or employers know about assistive technologies. And for some people, such technology doesn't yet go quite far enough (see 'Breaking down the barriers', previous page).

This may be reflected in the low number of disabled researchers. In 2000, the NSF estimated that 365,500 people with disabilities were employed in science, technology, engineering and mathematics (STEM) in the United States



Jesse Leaman (left) has devised an electronic 'rear-view mirror' for wheelchair users.

— about 7% of the science workforce. But there are more people with disabilities in the general workforce (13%), and even the college-educated workforce (9%). "There is a strong underrepresentation of persons with disabilities in STEM fields, especially at the PhD level," says Ted Conway, director of an NSF programme researching disabilities education.

Gardner is optimistic about the future, but sees change happening slowly. "I don't want to be too negative," he says. "I am optimistic about public awareness and new technologies. I believe that things are getting better, but it would be dishonest to say that things are already better."

One student who has persisted in the face of these challenges is Jesse Leaman. Despite being paralysed from the neck down ever since an accident when he was 18, Leaman has pursued a career in astronomy and graduated from the University of Maryland in College Park in 2002. During his time there, Leaman participated in a AAAS programme that places disabled students with an aptitude and passion for science into summer internships with research agencies or companies.

In 1998, Leaman spent the summer working in the microgravity department at NASA's Marshall Space Flight Center in Huntsville, Alabama, where he designed educational web pages about the working environment aboard the International Space Station. The next summer, he worked as an intern in the space science department at NASA's Goddard

Space Flight Center in Greenbelt, Maryland. Using voice-activation software called Dragon Naturally Speaking, he ran satellite data through an analysis program to study how the solar wind interacts with Earth's electric field.

It was during these internships that Leaman developed what he calls the Griffin Shield, a navigation and communication system for motorized wheelchairs and scooters. Leaman kept colliding with people and objects while manoeuvring in hallways and offices, so with the help of NASA technologists, he attached a small video camera to the rear of his wheelchair and an LCD monitor to the front. In effect, he now has an electronic 'rear-view mirror', powered by the chair's battery, which can also be hooked up to a laptop for communication purposes.

Leaman designed the Griffin Shield out of necessity and never imagined he would one day establish his own business. But when the shield turned out to be so practical and convenient, he says, "I knew I wanted to share it with other wheelchair users."

Next year, Leaman will receive his doctorate in astronomy from the University of California, Berkeley, and plans to continue collaborating with astronomers at Goddard. He is currently working on determining the rate at which supernovae occur.

He is also in the process of mass-producing the Griffin Shield and negotiating contracts with major wheelchair manufacturers. But unlike Gardner, Leaman has no intention of giving up his love for astronomy. "I can do both," he says. ■

Jessica Ebert is an intern in Nature's Washington DC office.

"This is about finding a way to make graphical information accessible to the blind and those with print disabilities. That is still one of the biggest barriers." — John Gardner

In a consulting room in Santa Barbara, California, Michael Salsbury found himself comforting his weeping doctor. It was July 2000, and he and his wife Gabriella were worried about their baby daughter, Gracie-Sophia, who was having seizures and losing her ability to stay alert and feed. The doctor had examined her and confirmed the Salsburys' worst fears: Gracie's brain was vanishing, and he couldn't tell them why.

The doctor's tears were born of frustration and despair. Gracie was not the first of the Salsburys' children to succumb to this mysterious illness. Two other girls had perished before their first birthdays from a strange, incurable disease that robbed them of mobility and caused parts of their brains to melt away. None of their doctors had seen anything like it, and for years, the Salsburys were unable to get a full diagnosis.

Now, thanks in part to their fortitude in donating their daughters' organs for research, the Salsburys have an answer. And it seems that the inherited disease that claimed Gracie and her sisters has implications far beyond the realm of rare neurological disorders. The faulty genes involved help control a vital aspect of our biology: the translation of RNA into protein. Biologists are realizing that documenting how this control can break down should help us to understand more about common diseases such as cancer and Alzheimer's — and may one day provide new ways to treat them.

For the Salsburys, finally getting a proper diagnosis for their lost children has helped them to heal. "We went as low as any family can probably get, three times," says Michael. "Now we finally know what took the lives of our daughters."

The Salsburys' troubles began in 1993, after the birth of their second child, Stephanie. At three months, she began to have subtle epileptic seizures. Her condition soon deteriorated, while doctors ran every possible test and tried different combinations of drugs. "She was getting everything around the clock," says Gabriella, who works as an intensive-care nurse for newborn babies. "It didn't help." Stephanie died aged eight months.

No answers

Gabriella gave birth to another daughter, Jennifer, in 1995. At first, Jennifer was alert and lively. But at three months, she began to develop the same ominous symptoms. This time, knowing that nothing could be done, the Salsburys refused most medical tests. "We had been through the wringer with Stephanie," says Michael. "We wanted to keep our baby at home and comfortable, and love her for as long as we could."

Shortly afterwards, Michael, a banker, was transferred to Switzerland, where they had a healthy daughter, and then Gracie-Sophia was born. This time, Gabriella sensed that something was wrong before the symptoms began. Swiss doctors dismissed her fears, and the

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

Lost in translation

A mysterious disease that causes children's brains to melt away is caused by errors in RNA translation. But biologists are realizing that this horrifying condition could shed light on more common problems. **Claire Ainsworth** reports.

family planned a baptismal celebration back home in Santa Barbara. But as the plane touched down, Gabriella saw that Gracie was having subtle seizures. "And so it became a funeral again," says Michael.

After Gracie died, the Salsburys abandoned plans for a bigger family. Believing their daughters were the only known cases of the illness, they donated the girls' organs to research and hoped that someone would find an answer.

But the Salsburys were not alone. From the early 1990s, several doctors around the world had been reporting bewildering cases of ailing or comatose children who had unusual brain scans¹. Raphael Schiffmann, a child neurologist at the National Institute of Neurological Disorders and Stroke in Bethesda, Maryland, was the first to describe and put a name to the disorder: CACH, for childhood ataxia with central hypomyelination².

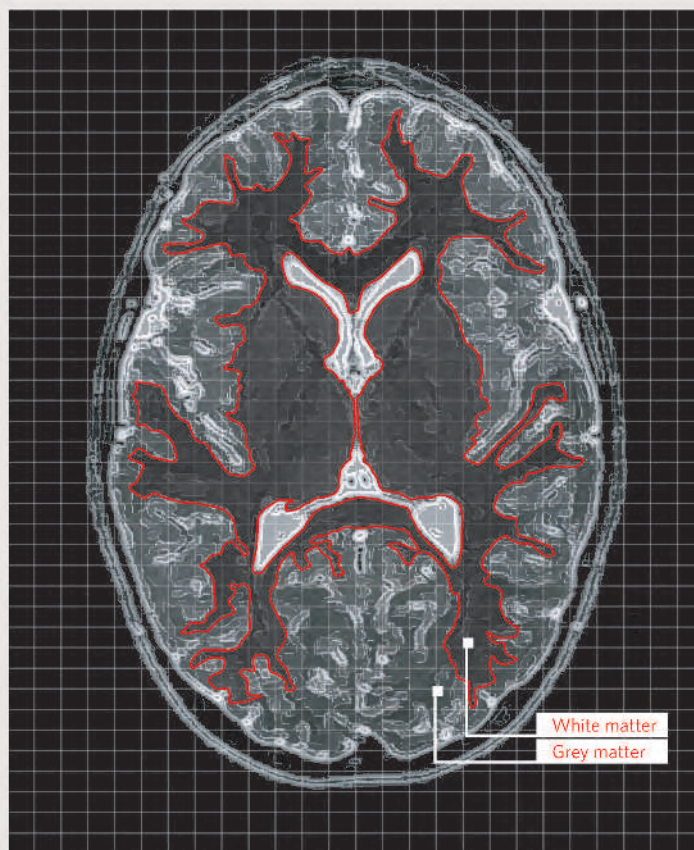
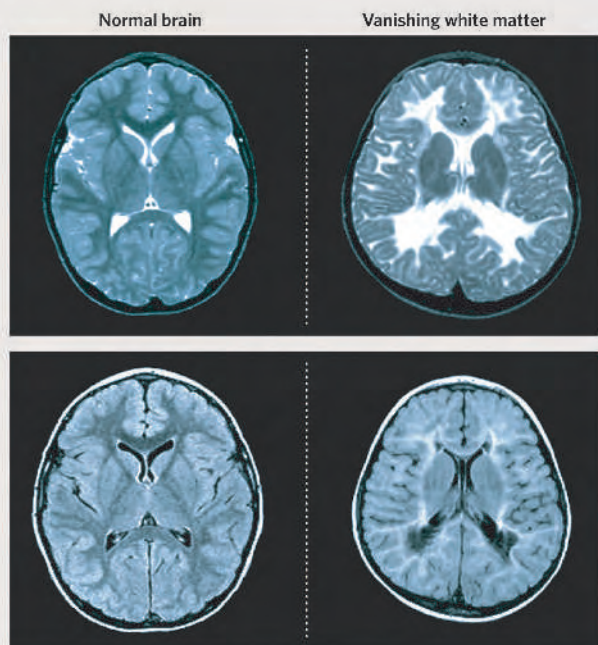
At the Free University Medical Center in Amsterdam, neurologist Marjo van der Knaap was also seeing similar cases. The children,

normally less than five years old, would initially seem normal. Then, following a bang on the head, fever or even a fright, they would begin to lose muscle coordination. In some cases, the subsequent decline came in episodes and took years. But in others, it was frighteningly rapid. "On, say, Thursday, a child is unwell — starting to have a fever or a little bit of a cold," explains van der Knaap. "Then on Friday, the child becomes irritable and limp, lies on the couch and doesn't do much. By Sunday, the child is in a coma."

Magnetic resonance imaging (MRI) scans showed that the white matter in the patients' brains was abnormal. White matter acts as the main communication link for sending commands from the brain's grey matter to the rest of the body. It consists mostly of axons, the long extensions of nerve cells that form the links to different parts of the nervous system. These connections are encased in a substance called myelin, which acts like the plastic insulation around an electrical wire. With the loss

MYSTERIOUS DISAPPEARANCE

Two types of magnetic resonance image highlight the devastating effect of a genetic condition that destroys the brain's white matter (outlined in a normal brain, right). In the upper pair, the white matter (dark grey) in the normal brain is largely replaced by cerebrospinal fluid (white) in the diseased organ. Similarly, in the second images white matter (grey) is replaced by fluid (dark grey).



M. S. VANDER KNAAP/VUUNIV. MED. CENTER

of white matter, instructions to the body would fail to get through. This fits with the symptoms: patients' cognitive function is largely unaffected, but their motor skills can be destroyed.

The MRI scans were clear: the white matter was melting away and being replaced with cerebrospinal fluid, the watery liquid that bathes the brain and spinal cord³. This led van der Knaap to name the condition 'vanishing white matter' or VWM disease — now known as VWM/CACH. Schiffmann has a theory about how the disease progresses. "The first thing that goes wrong is the insulation of the wire," he says. "Then at some point, the wires themselves begin to tear." In the end, the white matter disappears completely.

By the late 1990s, the hunt for the genetic roots of VWM/CACH was on. We now know that it can be caused by mutations in five different genes, each on a different chromosome^{4,5}. Normally, this scattering would have made it all but impossible to track the genes down. But a quirk of Dutch genealogy came to the rescue.

Local knowledge

Van der Knaap's team was collecting samples and data from families in the Netherlands, where 1 in 40,000 people are affected. Most of the families involved came from a rural region in the east of the country where the population has tended not to migrate or intermarry with people from elsewhere (see right). This gave a relatively homogeneous genetic background

against which it was easier to detect disease-causing genes. And the researchers had a further stroke of luck when they found that several affected families shared a single common ancestor. Within this 'superfamily', the researchers were able to scan the genomes of affected and unaffected individuals to identify a gene involved in VWM/CACH.

The culprit, the gene for part of a protein called eIF2B, came as a surprise⁶. Far from being specific to the brain, this protein is present in almost every cell in the body. It plays a key role in stitching together amino acids to make proteins according to the recipe laid down in RNA, a process known as translation. The eIF2B protein is made of five fragments or subunits, coded for by the five different genes. It unites with ten other proteins to form a 'machine' that helps control the translation process. And when van der Knaap and her colleagues looked at other VWM/CACH patients, they found that they had mutations in the genes for these other subunits^{4,5}.

The discovery that faults in the control of translation can cause such a dramatic disease has thrown a fresh spotlight on the importance of the process in health and disease⁷. "It's triggered wider excitement," says Christopher Proud, now at the University of British Columbia in Vancouver, Canada, who has worked on eIF2B and related translation factors for some 15 years.

When thinking about the control of the production line from gene to protein, most

HUNTING THE GENE



In the eastern part of the Netherlands, people tend to live in small villages and do not often relocate. This allowed researchers to track down the faulty gene. They identified 252 ancestors clustered in the area of Zwolle. The numbers indicate the percentage of ancestors living in each village.

interest has focused on the earlier process of transcription. In this step, DNA is 'read' to produce messenger RNA, or mRNA. Various proteins known as transcription factors control this process, and they have come under intense scrutiny from cell biologists and researchers interested in diseases such as cancer.

The mRNA gets carefully edited and is then sent to complex molecular 'robots' called ribosomes, which read it and slot the amino acids together in the order detailed in the code. Together with its partners, eIF2B helps to position the ribosome at the correct starting point as it scans along the mRNA.

One intriguing aspect of VWM/CACH is that its severity can vary, depending on where the mutation lies in the eIF2B protein. Proud and his colleagues suggest that this is because different mutations have different effects on eIF2B's ability to function and interact with the other proteins that regulate its activity⁸. It seems that different mRNAs have different sensitivities to the level of eIF2B. This might help explain the puzzle of why the brain, and particularly the cells that make myelin, is most affected in VWM/CACH. If these cells rely especially heavily on eIF2B to make their key proteins, they will be more sensitive to drops in its level of activity.

Communication breakdown

The emerging understanding of VWM/CACH fits with a growing realization that the control of translation is more subtle than anyone thought. "It took a long time, but the wind is changing," says Arrigo De Benedetti, a molecular biologist who works on translation at Louisiana State University in Shreveport. "People are finally recognizing protein synthesis as a major player in disease in general, and in cancer in particular."

The suggestion that faulty translation could cause cancer has been around for about 15 years. In 1990, a team led by Nahum Sonenberg at McGill University in Montreal, Canada, showed that another translation factor, called eIF4E, can make cells become cancerous if present in abnormally high amounts⁹. In patients, high levels of this protein are also associated with resistance to anticancer therapy and with the disease's spread through the body. Meanwhile, other cancer researchers have noticed that some proteins are present at abnormal levels in cancer cells despite there being no obvious problem with the production of their mRNAs — hinting that translation, not transcription, is at fault.

Some mRNAs respond more to changes in the activity of eIF4E than others. Biologists now know that this is because each mRNA has a long, untranslated region at its start that affects how it responds to different translation factors, depending on its sequence or three-dimensional structure¹⁰.

As if this weren't complicated enough, there is another kind of untranslated region that also affects how mRNAs respond to translation



Gabriella and Michael Salsbury lost three of their baby daughters to a rare brain disease.

factors. These regions are long — up to a third of the length of the entire mRNA. They fold into elaborate structures, called internal ribosome entry sites, or IRESs, that together with control proteins summon ribosomes to read the mRNA¹¹. Mutations in IRESs have been linked to a number of diseases. They have been found in cancer patients in genes that control cell division, for example.

Other translation-control mechanisms are probably waiting to be discovered. "There could be lots of other things out there," says Anne Willis, a molecular biologist at the University of Nottingham, UK, who works on IRESs and their role in cancer. "They've just not really been looked for."

Unravelling the complexities of translational control may look daunting, but its very subtlety may be to clinicians' advantage. In theory, it should be possible to selectively target the production of particular proteins implicated in disease without causing wide-ranging side effects. "There are probably dozens, if not hundreds, of compounds that can inhibit steps in translation and have not been exploited yet to their full extent," says De Benedetti.

Targeted effect

In fact, one such drug is currently in clinical trials to treat cancer. Rapamycin, originally developed as an immunosuppressant, targets one of the systems that regulates translation¹². The drug shuts down a protein called mTOR, which senses aspects of the cell's environment, such as the availability of nutrients, and allows the cell to grow and divide if all is well. One of the ways it does this is by controlling the activity of translation factors. In many cancers, normal mTOR signalling seems to break down.

It should also be possible to target specific translation factors directly. And in the long run, Schiffmann and others hope that it may be possible to help VWM/CACH patients by targeting eIF2B or the affected mRNAs. "The problem is that it is really too early to say how one would do that," says Proud. "We need to know a lot more about how the condition works."

Researchers working on VWM/CACH are now concentrating on developing mouse models of the disease and studying the biology

of affected brain cells. They are also identifying other variants of VWM/CACH caused by different eIF2B mutations^{13–15}. For example, van der Knaap has identified mutations that cause severe, early-onset disease that affects multiple organs and causes babies to die before birth or in the first few months of life¹⁶. It was while working on that project that she happened on organ samples and MRI data from a family that had lost three baby girls to an undiagnosed white-matter disease.

So it was in 2003 that the Salsburys got the phone call they had long waited for. It was van der Knaap, who had discovered eIF2B mutations in their daughters' brain tissue. As well as helping the Salsburys recover from their loss, the knowledge offers the family's surviving children the opportunity to be tested for the mutations, so they won't have to endure the same ordeal when they come to have families of their own.

The Salsburys have now set up a charity, the Three Little Angels Foundation, to raise money for research and to improve awareness of rare neurological diseases. They are passionate advocates of donating organs for research. "We see science and medicine as an evolving effort," says Michael. "Each single daughter gave science a new piece of knowledge."

Claire Ainsworth is a senior news & features editor for Nature.

1. Hanefeld, F. *et al. Neuropediatrics* **24**, 244–248 (1993).
2. Schiffmann, R. *et al. Ann. Neurol.* **35**, 331–340 (1994).
3. van der Knaap, M. S. *et al. Neurology* **48**, 845–855 (1997).
4. Leegwater, P. A., Pronk, J. C. & van der Knaap, M. S. *J. Child Neurol.* **18**, 639–645 (2003).
5. van der Knaap, M. S. *et al. Ann. Neurol.* **51**, 264–270 (2002).
6. Leegwater, P. A. J. *et al. Nature Genet.* **29**, 383–388 (2001).
7. Abbott, C. M. & Proud, C. G. *Trends Biochem. Sci.* **29**, 25–31 (2004).
8. Li, W., Wang, X., van der Knaap, M. S. & Proud, C. G. *Mol. Cell. Biol.* **24**, 3295–3306 (2004).
9. Lazaris-Karatzas, A., Montine, K. S. & Sonenberg, N. *Nature* **345**, 544–547 (1990).
10. De Benedetti, A. & Graff, J. R. *Oncogene* **23**, 3189–3199 (2004).
11. Stoneley, M. & Willis, A. E. *Oncogene* **23**, 3200–3207 (2004).
12. Bjornsti, M.-A. & Houghton, P. J. *Nature Rev. Cancer* **4**, 335–348 (2004).
13. Fogli, A. *et al. Am. J. Hum. Genet.* **72**, 1544–1550 (2003).
14. Fogli, A. *et al. BMC Women's Health* **4**, doi:10.1186/1472-6874-4-8 (2004).
15. Fogli, A. *et al. Ann. Neurol.* **52**, 506–510 (2002).
16. van der Knaap, M. S. *et al. Am. J. Hum. Genet.* **73**, 1199–1207 (2003).

BUSINESS

More flavour up front

Two young biotechnology start-ups in the United States are leading an innovative search for powerful flavour enhancers. If they succeed, biologists at the companies say, they could revolutionize food manufacture by dramatically shrinking the amounts of sugar, salt and other flavourings used by the industry.

Senomyx of La Jolla, California, and New Jersey-based Linguagen are using high-throughput techniques pioneered by the drug industry to scan thousands of compounds for their impact on human taste receptors. These poke out of tongue cells and bind to food molecules, allowing us to taste salt, sweet, bitter, sour and umami, the flavour of monosodium glutamate.

Armed with bulk screening and genetic-sequence information for the receptors and their associated pathways, the companies are trying to pinpoint and refine the natural and synthetic compounds that can do the job.

"The way an enhancer works is that it makes the taste receptor more efficient," says Mark Zoller, head of research at Senomyx, the larger of the two companies, which raised \$34 million in an initial public offering on NASDAQ last June. Originally called Ambryx, Senomyx was founded in 1998 by Charles Zuker, a molecular geneticist at the University of California, San Diego, who had discovered and described some of the taste receptors.

Linguagen founder Robert Margolskee, a biophysicist at the Mount Sinai School of Medicine in New York, set up his company after describing some of the other steps in the molecular pathway between food hitting the tongue and the brain recognizing a taste. Both companies really took off after the Human Genome Project made it much easier to identify the genes behind taste and express them in cells for screening.

In a typical screening process, active genes for the receptor or pathway element are inserted into mammalian cells, and the resulting transgenic cells are grown in 384-well plates. A slightly different compound is added to each well from a small-molecule library. If the com-



BANANASTOCK/ALAMY

Sweet and sour: flavour enhancers would allow the food industry to cut back on sugar and salt.

pound binds or flirts with binding with the receptor, the activity will make the dye in the well light up. That compound will then be plucked out and its exact effect on the taste process investigated.

Because these compounds target the receptor so well, they are effective in tiny quantities — small enough to qualify for an expedited safety approval process, which should take three years or less.

Senomyx, whose stock has doubled in value since last year's flotation, plans to make its money by licensing the intellectual-property rights to food companies. It already has research agreements with a brace of food giants, including Nestlé and Coca-Cola.

The two start-ups hold patents on the specific screening processes, as well as compounds they have identified as promising — advantages that could aid their survival. Investors are betting that Senomyx, at least, will turn a profit one of these days. "If things continue to go well, there is the chance that this could become a very big company," says David Weber, a stock analyst at First Albany in New York state.

Senomyx is also sitting on intellectual property relating to the olfactory receptors. Although it hasn't started an active research programme with the nose's machinery for recognizing smells, it is thinking about it.

Clint Brooks, head of research at New York-based International Flavor and Fragrance, points out that this is a key consideration: the five fundamental tastes are just the base notes to flavours, which also involve many of the 300 or so olfactory receptors. He adds that while companies such as Linguagen are discovering novel, exciting molecules, these must be incorporated into the overall flavours of foods through the "art and science" of an experienced flavourist. ■

Emma Maris

On the up: Senomyx stock



IN BRIEF

BLUE CHIPS IBM has announced a tie-up with chip-makers in Taiwan and Germany to jointly research computer memory chips based on phase-change materials.

The materials, whose structures change from crystalline to amorphous when an electric current passes through them, have already attracted the interest of researchers at companies such as Philips in the Netherlands (M. H. R. Lankhorst *et al.* *Nature Mater.* 4, 347-352; 2005).

Macronix of Hsinchu and the Munich-based Infineon will now join in a project involving 20 to 25 researchers working full-time at IBM's research labs in San Jose, California, and Yorktown Heights, New York.

STEM-CELL START California biotechnology firm BioDefense, which specializes in homeland-security work, says that it is planning an initial public offering of stock to raise funds for its stem-cell division.

The company adds that its newly founded Stem Cell Research division would specialize in the possible use of embryonic stem cells to treat acute liver disease.

Chief executive David Chin says he believes that stem cells would "allow for the development of an artificial liver". But the company has given no schedule for its stock offering.

NO GREEN LIGHT ExxonMobil's shareholders rejected a motion on climate change at the company's annual meeting on 25 May — although holders of almost 30% of the stock voted for it.

The measure would have mandated a report on how the company planned to deal with the regulations of the Kyoto Protocol in countries where it is in effect, and a study of the feasibility of reducing emissions in the United States.

This was the latest in a series of similar efforts by shareholders at major corporations (see *Nature* 435, 410-411; 2005).

Technology managers do their bit for world health

SIR — Your Editorial “Wanted: social entrepreneurs” (*Nature* **434**, 941; 2005) made several disparaging remarks about the academic technology-transfer profession to which I wish to respond.

You stated: “University technology offices tend to patent aggressively, look no further than generating income, and often fail to include provisions beneficial to tackling orphan diseases in their licensing deals with companies.”

First, the top priority of any office of technology transfer is to get a qualified company to make a serious commitment to develop each technology. Of course we negotiate hard to ensure that if the project is a success we are fairly compensated, but the most contentious part of most negotiations is generally not the financial terms but the due-diligence terms — the resource commitment that the company makes to develop the technology.

As Mary Sue Coleman, president of the University of Michigan, recently said in

remarks to the 2005 Annual Meeting of the Association of University Technology Managers (AUTM): “Our interest in commercializing technology, in nurturing start-up companies and facilitating patents and licence agreements is not about the promise of future revenues. Of course, revenue generation serves as an incentive. But first and foremost, tech transfer must serve our core mission: sharing ideas and innovations in the service of society’s well-being.”

Second, I disagree with your assertion that licensing offices are not sensitive to global health issues. In the AUTM 2003 Annual Survey, one of the transactions showcased was a new treatment for Chagas’ disease discovered by Washington and Yale universities and licensed by them to the Institute for OneWorld Health, the company that was the focus of your Editorial.

Third, the contact you advocate with the Centre for the Management of Intellectual Property in Health Research and

Development (MIHR) is already under way. For several years there has been a positive dialogue between AUTM and MIHR, largely driven by individuals who belong to both organizations. The focus of the 2006 Annual Meeting of AUTM is improving society, and the plenary session will focus on global health and the elimination of intellectual-property barriers to bringing advances in neglected diseases to the developing world.

The biggest challenge will be to get the pharmaceutical and biotechnology industries — whose involvement in translating academic discoveries into safe and effective treatments is critical — to accept our proposals. But at least universities can provide leadership and start the discussion.

Ashley J. Stevens

Office of Technology Transfer,
Boston University, Boston,
Massachusetts,
and Association of University Technology
Managers, Northbrook,
Illinois 60062, USA

Ampicillin threat leads to wider transgene concern

SIR — We are concerned by the suggestion, in your Editorial “Don’t rely on Uncle Sam” (*Nature* **434**, 807; 2005), that the US Food and Drug Administration does not consider the presence of the ampicillin-resistance gene in Syngenta’s unapproved variety of genetically modified *Bt10* maize to represent a safety problem.

This is not the view of the UK government’s scientific advisers (the DEFRA Antimicrobial Resistance Coordination Group), who state that some important veterinary pathogens remain susceptible to ampicillin (K. L. Goodyear *et al.* *J. Antimicrob. Chemother.* **54**, 959; 2004). They state that there is “extremely low or no detected resistance in certain bacterial species”, so that “any occasional transfer of resistance genes to these organisms would be a very significant event”. If, as a result of such horizontal gene transfer, it became necessary to use more modern antimicrobials to treat animal disease, they write, “then there could be significant consequences for the consumer through the food chain”.

The risk of horizontal gene transfer from genetically modified organisms (GMOs) is not a theoretical one. One study found that, after *Bt* genes in plasmid form were incubated in the saliva in a sheep’s mouth for a few minutes, they could still transform *Escherichia coli* bacteria so that they developed antibiotic resistance (P. S. Duggan *et al.* *Br. J. Nutr.* **89**, 159–166; 2003).

In addition, it is worth noting that the ampicillin-resistance gene in *Bt10* maize and other genetically modified crops is a remnant of the bacterial plasmid inserted into these varieties, and would therefore function very efficiently if taken up by bacteria as a result of horizontal gene transfer.

Once the *Bt10* maize incident has been dealt with, we feel there should be a review of the general question of horizontal gene transfer from GMOs. There is no reason to believe that any health implications are confined to antibiotic-resistance marker genes; they could, for example, equally apply to the inserted *Bt* toxin genes present in all genetically modified *Bt* crops. However, the transfer of antibiotic resistance is the only such risk currently being addressed by the authorities that regulate GMOs.

We consider that the case-by-case approval approach used by the authorities does not adequately address such problems, which are common to all GMOs.

Gundula Azeez

Soil Association, 40–56 Victoria Street,
Bristol BS1 6BY, UK

Activists should accept mainstream view of GM

SIR — It is gratifying to read, on your Correspondence page, that environmental campaigners are urging the public to accept the view of a consensus of climatologists, glaciologists and atmospheric physicists that “anthropogenic climate change is a reality”

“An overwhelming majority of plant geneticists, biochemists and molecular biologists have endorsed the use and safety of genetically modified crops.”

(“Time to speak up for climate-change science” *Nature* **434**, 559; 2005). Having accepted the expertise of scientists on this issue, perhaps Greenpeace and Friends of the Earth should reconsider their opposition to genetically modified (GM) crops, as an overwhelming majority of plant geneticists, biochemists and molecular biologists have endorsed the use and safety of these crops. This would allow the economic, environmental and humanitarian benefits of this technology to be fully realized.

As president of a biotechnology company and emeritus professor of biology at Queen’s University, Ontario, I agree with the environmentalists that scientists should make their science fully accessible to the general public. If this had been done, all the problems of misinformation and concern about GM use and safety would have been avoided.

David T. Dennis

President and chief executive, Performance
Plants, BioScience Complex, 116 Barrie Street,
Kingston, Ontario K7L 3N6, Canada

Contributions to Correspondence may be submitted to corres@nature.com. They should be no longer than 500 words, and ideally shorter. Published contributions are edited.

COMMENTARY

A fresh start for Europe's space agency

The European Space Agency has a strong track record and plenty of ambition to propel it into its next 30 years, says **Giovanni Bignami**. But key decisions must be made in the context of a new Europe.

Celebrations for the thirtieth birthday of the European Space Agency (ESA) began early when the Huygens probe landed on Saturn's moon Titan on 14 January. Another mark of European success in space science is that out of a fleet of 17 satellites operated by ESA, 15 are involved in scientific projects. A further satellite will soon be launched in the first European mission to Venus.

Of course, Europe has other space successes to celebrate: the 1975 COS-B gamma-ray telescope, with which ESA began; the Giotto mission sent to meet Halley's comet; the Hipparcos satellite that mapped the heavens; and much else besides¹. The Ariane launcher, French-driven but solidly European, now dominates the world market in commercial satellite launches, and is a long way from Veronique, Blue Streak and Europa, the forgotten names of the first stages of European rocketry.

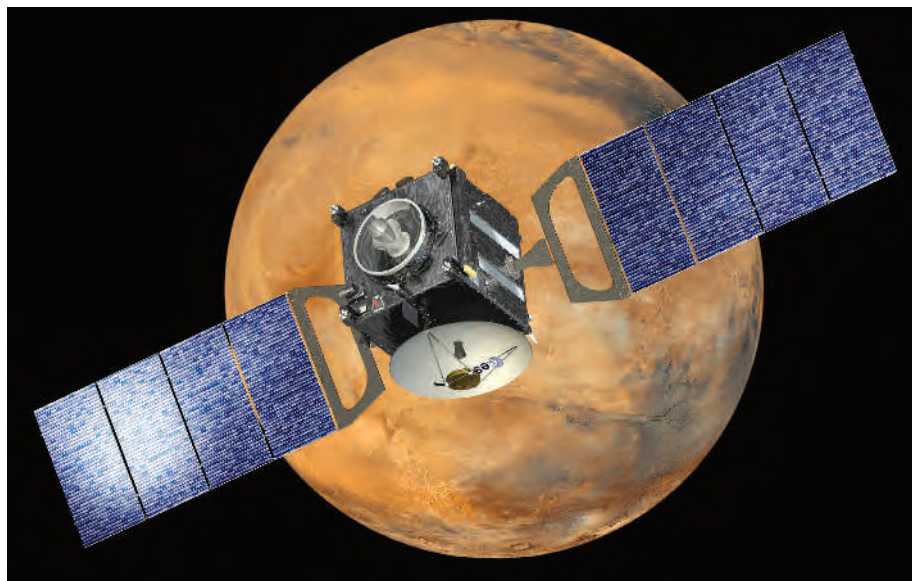
Today, ESA has a unique opportunity to explore the Solar System, especially if it can quickly shift its attentions beyond the International Space Station. And now is a good time for Europe to take the lead in a number of scientific endeavours, mostly in astronomy, that have been left open by NASA's change of direction towards exploration. Ambitions also reach well beyond space science: the Galileo programme for global positioning, and the Global Monitoring for Environment and Security Initiative are both due to be established by the end of 2008.

Rallying support

But continued success will need continued investment. The United States has long understood the importance of investing in basic science in space. Despite the US gross national product being similar to Europe's, NASA's budget is more than five times ESA's, which today stands at €3 billion² (US \$4 billion). Moreover, NASA's science programme accounts for 30% of its overall budget — rich by comparison with ESA's 12%.

Of the €1 million spent by ESA every day on science, more than 80% goes to industrial contracts. These are particularly cherished by Europe's aerospace industry; scientific projects generate industrial competitiveness and innovation³.

As the European Union (EU) expands, ESA's membership grows; this year, it includes



STONE/GETTY IMAGES

Missions to Mars: building on the success of Mars Express, European scientists have set their sights on future missions to the red planet, to collect more clues about its atmosphere and surface.

17 nations. With the EU's science budget set to double over the next five years, we, Europe's space scientists, are conscious that the importance of our job goes beyond our community. But we also know we have to work with limited financial support. We are used to it: we don't complain, we get organized.

From 1984 to 2004, ESA science was guided by two ten-year plans; a third will take us to 2014. This third plan includes numerous missions whose aims range from studying Mercury to understanding the frontiers of cosmology. Like Soviet-style planners of yore, we are now building our fourth ten-year plan — Cosmic Vision 2015–2025. But unlike those of Soviet planners, this plan is built according to Abraham Lincoln's doctrine, “by the people, for the people”.

In reaction to a request issued last year, 151 European groups entrusted their ideas to ESA. This is more than twice the number of groups that have responded previously, and indicates the increased number and diversity of European space scientists. Before the year is out, our Cosmic Vision plan, already discussed with the space-science community and approved by ESA's advisory structure, will be presented to ESA's council of ministers, who dictate the agency's financial portfolio. ESA's

resources cover the ‘mandatory science programme’ as well as several optional programmes that range from Earth observation and telecommunications to the Space Station and Aurora (see below). For the mandatory programme, contributions from ESA-member states are made according to their gross national products; for the optional programmes, they are made à la carte.

Pooling efforts

What is in our 2015–2025 plan? The Cosmic Vision plan that was released in April is organized around four grand questions. These range from ‘what are the conditions for planetary formation and the emergence of life?’ to ‘how did the early Universe originate and what is it made of?’ Each of these questions prioritizes specific subjects for study, ranging from life and habitability in the Solar System to the evolving violent Universe. Starting from these objectives, strategies have been drawn up for possible space missions in astronomy, fundamental physics and solar-system sciences.

Cosmic Vision also includes a technology development plan that spells out how to get from these strategies to proto-missions — the concepts from which real missions will emerge. Examples of proto-missions include:



COS-B, one of Europe's most successful missions, and ESA's first, was used to study gamma-ray sources.

an X-ray Super Telescope; a next-generation gravity-wave explorer; and a Jupiter/Europa probe. Although ministers at December's council meeting will not be allocating funds to the Cosmic Vision plan, a broad commitment is hoped for. If all goes well, calls to the community for concrete mission proposals will start early next year.

But the European space-science programme cannot proceed in isolation, without input or overlap with other ESA programmes, such as Aurora and the Space Station.

Aurora was conceived in 2001 as a pay-as-you-go programme to explore the Solar System, and to support European aerospace. It was not originally conceived with the main purpose of doing science, and must remain true to its original mission.

Because it relies on voluntary contributions, Aurora has had a slow start. The biggest contributors so far are Italy and the United Kingdom; exact sums are now being discussed. These contributions may be matched by Germany, despite its difficulties with the Space Station (see below). France has a self-imposed cap on its overall contribution to ESA; any extra money invested into Aurora means less money spent on other ESA projects. But overall, the €100 million needed to start a serious effort seems to be within reach.

Aurora was intended to provide a way forward for Europe's aerospace industry, after the (assumed) completion of the Space Station. But progress with the Space Station has met with severe difficulties, generated by the Columbia tragedy and the subsequent hiatus in Shuttle flights: no flights have taken place

for more than two years. Return to flight is expected next month (fingers crossed), but there are rumours that the number of remaining flights will be as few as 15, or half the number previously considered.

This drop in flight number suggests a lack of enthusiasm on the part of NASA for even a half-decent scientific use of the Space Station, despite the agency's earlier claims that science was the Station's *raison d'être*. It also raises questions about the value of the investment put into it by NASA, Russia, Japan, many ESA states and others.

Worse still, the slumbering Space Station is holding back other space programmes, including Aurora. For example, while the German-led Columbus module for the Space Station — so named because it was meant to be in orbit in 1992 — remains on the ground in Bremen, the German delegation has hesitated to enter the Aurora programme. A quick way forward must be found in collaboration with the new NASA, once the shuttle returns to flight.

Aurora uplifted

A potential silver lining is that any quick resolution of the Space Station problem could release new resources for a well-focused Aurora programme. ESA has wisely created a new directorate that is dedicated to both robotic exploration and human presence in space. Its funding covers both Aurora and the Space Station. So money not going into the Station could easily be transferred to Aurora, if member states agree.

At the upcoming ESA Council, we must make ministers an offer they cannot refuse, and

one that does not seem to compete with the mandatory science programme. We should clearly distinguish the technology and exploration content of Aurora from the aims of the science programme, as outlined in the Cosmic Vision plan. Here are some suggestions.

As soon as possible, Aurora should supersede the Space Station as a programme that drives industry, while also satisfying taxpayers' cosmic dreams. Aurora's aim should be to use innovative technologies to explore the Solar System — with a long-term goal of putting people on Mars. New methods need to be found for nuclear-power generation and propulsion, but also for landing on Mars, such as aerocapture and aerobraking (to aid spacecraft landing in the thin martian atmosphere). The development of such techniques would revolutionize exploration and allow for more efficient payload transportation.

Furthermore, when planning robotic exploration of Mars, mission and payload selection should be delegated to the science directorate — well versed in the time-honoured peer-review approach. The same directorate could even contribute to the allocation of funding for scientific payloads.

We should also start with something credible. The Beagle 2 failure has put pressure on us to deliver. One concrete proposal emerged from an ESA workshop in Birmingham in April, in which scientists recommended starting with a rover mission to Mars that could eventually culminate in an international sample return mission. Finally, Aurora should embrace international partners but will need to maintain its own personality and independence, for example in developing key enabling technologies, such as aerocapture.

President George Bush's new Moon–Mars plan for NASA may leave more room for ESA in space science. It certainly seems the right moment for a stronger Europe–United States partnership, at least in some fields such as X-ray astronomy, or outer-planet exploration. A grave difficulty here is caused by the current US restrictions on technology exchange. Such restrictions hamper joint work and are especially hard on US industry.

ESA has other partners as well. Russian scientists are working on the gamma-ray astronomy observatory Integral. And now China is involved in the Double Star mission for studying the Earth–Sun connection. The Bepi-Colombo joint mission with Japan to Mercury is already in the making. And a very recent agreement has been reached with India, for ESA's participation in India's lunar mission.

The opportunity to give Europe a central, aggregating role in space science is now here; it augurs well for ESA's next 30 years. ■

Giovanni Bignami is at the Centre d'Etude Spatiale des Rayonnements, 9 avenue du Colonel Roche, 31028 Toulouse, France.

1. <http://www.esa.int/esaSC/index.html>
2. Andersen, B. *Science* **307**, 1206 (2005).
3. Coppens, Y. et al. *Le Monde* 4 April, 1 (2005).

BOOKS & ARTS

Expanding evolution

A broader view of inheritance puts pressure on the neo-darwinian synthesis.

Evolution in Four Dimensions: Genetic, Epigenetic, Behavioral, and Symbolic Variation in the History of Life

by Eva Jablonka & Marion J. Lamb

Bradford Books: 2005. 462 pp.

\$34.95, £22.95

Massimo Pigliucci

There have been rumblings for some time to the effect that the neo-darwinian synthesis of the early twentieth century is incomplete and due for a major revision. In the past decade, several authors have written books to articulate this feeling and to begin the move towards a second synthesis. David Rollo, in his book *Phenotypes* (Kluwer, 1994), was among the first to attempt to bring the focus back to the problems posed by phenotypic evolution. In *Phenotypic Evolution* (Sinauer, 1998), Carl Schlichting and I framed the debate in terms of the integration of development, environment and genetics by articulating the concept of “developmental reaction norms”. Stephen Jay Gould then produced an overly long (and at times acrimonious) sketch of the new synthesis in *The Structure of Evolutionary Theory* (Harvard University Press, 2002). Finally, Mary-Jane West-Eberhard, in *Developmental Plasticity and Evolution* (Oxford University Press, 2003), greatly expanded on my book and the one by Rollo, producing the most comprehensive alternative account of evolutionary theory yet. *Evolution in Four Dimensions* by Eva Jablonka and Marion Lamb is the most recent addition to this genre, and contributes yet another valuable perspective to the discussion.

Jablonka and Lamb provide a framework that includes not one but four sources of inheritance in living organisms: there is the standard genetic one, based on nucleic acids such as DNA and RNA; there are epigenetic inheritance systems, such as (but not limited to) chromatin marking systems and RNA-interference systems for gene silencing; third, there are behavioural inheritance systems, including behaviour-influencing substances (think pheromones) and social learning (both imitative and not); finally, humans have also developed a symbolic inheritance

system based on the ability to communicate by manipulating symbols.

The authors argue that there is more to heredity than genes; that some hereditary variations are non-random in origin; that some acquired information is inherited; and that evolutionary change can result from ‘instruction’ as well as selection. This may sound rather revolutionary, even preposterously close to lamarckism. But Jablonka and Lamb build on evidence from standard research in evolutionary and molecular biology, and their case should be examined on its merits, rather than being dismissed by a knee-jerk reaction.

Consider the charge of lamarckism. Jablonka and Lamb happily embrace the term, but with one important qualification. As they correctly point out, there are at least two very distinct meanings of the word. Most biologists associate lamarckism with the idea of direct adaptive feedback from the soma to the germ line. That version of lamarckism is dead, killed off by our understanding of

molecular biology, and nobody is attempting to revive it.

The second meaning is actually closer to the core of Lamarck’s ideas, which are rarely, if ever, read by modern biologists. The suggestion is that some heritable, adaptive changes come not from natural selection, but from the action of evolved internal systems that generate non-random ‘guesses’ in response to environmental challenges. Examples are not hard to find, contrary to the assumed wisdom of standard neo-darwinism. Consider the existence of ‘hotspots’ that make mutations in certain regions of the genome much more likely than in others. Or the impressive ability of some bacteria to increase the mutation rate of a specific gene involved in the metabolism of a given amino acid when that amino acid becomes scarce in the environment.

Jablonka and Lamb are surely taking a gamble in labelling their position as lamarckist, but they are correct to point out that no modern biologist is a darwinist in the sense Darwin

would have understood — not least because Darwin included a lamarckian mechanism of the first (now frowned upon) type in his theory, as he had no solution to the problem of heredity.

If one accepts this bold, expanded version of heredity and evolution, it turns out that evolution can proceed very rapidly and phenotypic modification can precede genetic changes — something also suggested by several of the authors of the other books mentioned above. Indeed, changes at the genetic level will often simply stabilize adaptive modifications that are initiated through phenotypic plasticity, epigenetic control mechanisms, or behavioural and symbolic means. This is a framework that would greatly help to solve old problems in evolutionary biology, such as the origin of novel structures, and even the appearance of what ‘intelligent design’ proponents refer to, rather nonsensically, as ‘irreducible complexity’. This wouldn’t require the abandonment of neo-darwinism, but rather its expansion beyond what Ernst Mayr contemptuously labelled ‘bean-bag genetics’.

The irony, as Jablonka and Lamb point out, is that empirical evidence for

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

Down the generations: symbolic systems such as the written word provide an important means of cultural inheritance.

J. VINK/MAGNUM

the importance of epigenetic inheritance systems comes from the partial failure of the originally ultra-reductionist, gene-centred approach that gave us genomics. It is becoming increasingly clear that the interesting stuff is going on at the level of large gene networks, not of individual genes, partly because there is widespread functional redundancy in the genome. This is why we are seeing an astounding proliferation of 'omics' — after genomics, we have had proteomics, metabolomics and even phenomics, whatever that may mean.

Evolution in Four Dimensions also features a series of fictitious dialogues between the authors and a character named Ifcha Mistabra, which is Aramaic for "the opposite conjecture". This is a time-honoured philosophical device (used in the platonic dialogues, and in David Hume's dialogues on natural religion) for considering possible objections to one's arguments and discussing them in a literary way. Some

scientists may feel alienated by this device, but I found it refreshing to read a science book that is a conscious attempt at good literature. I really don't understand why so many of my colleagues equate boredom with seriousness.

The clamour to revise neo-darwinism is becoming so loud that hopefully most practising evolutionary biologists will begin to pay attention. It has been said that science often makes progress not because people change their minds, but because the old ones die off and the new generation is more open to novel ideas. I therefore recommend this and the other books I mentioned on the future of evolutionary theory to the current crop of graduate students, postdocs and young assistant professors. They'll know what to do. ■

Massimo Pigliucci is in the Department of Ecology and Evolution, State University of New York at Stony Brook, 650 Life Science Building, Stony Brook, New York 11794, USA.

that would surely be less damaging than losing half of all insect species with their pollinating services: our agricultural crops could be in trouble within short order. We live in a bug-driven world.

All this is dealt with in splendid detail in this book by Michael Samways, a leading entomologist at Stellenbosch University in South Africa. He starts out with the rationales for insect conservation, then considers such esoteric factors as evolutionary radiation, flight mechanisms, polymorphisms and taxonomic challenges. He reviews insects' roles as key-stone organisms, soil modifiers, pollinators, parasitoids and predators, and disease vectors. He considers insect survival in a fast-changing world, assessing such issues as environmental contamination, agricultural encroachment, deforestation, threats from invasive aliens, biological controls, genetic engineering, climate change and future evolution, as well as synergized interactions between these factors. The book concludes with an extended evaluation of conservation strategies, including reserve selection, plant and animal surrogates, phylogenetic considerations, inventorying and monitoring, species restoration, triage conservation (focusing efforts on the top priorities), and biodiversity hotspots as applied to insects.

Samways displays a flair for engaging asides, such as his comment on insects' fecundity: "One gravid aphid, left to reproduce with zero mortality, will, after one year, cover the globe with an aphid layer over 140 km thick."

There are very few insect books of such expansive scope, and this one could be a standard text for years. It will be welcomed by specialists in entomology, biodiversity, mass extinction, evolution and half-a-dozen associated fields. But it is much more than an expert book for experts; it should appeal to everyone interested in the fast-diminishing biodiversity of our planet. All in all, this is an expensive book that is excellent value. ■

Norman Myers is honorary visiting fellow at Green College, Oxford University, Upper Meadow, Quarry Road, Oxford OX3 8FS, UK.

Death and taxas

Insect Diversity Conservation

by Michael J. Samways

Cambridge University Press: 2005. 342 pp.
£30, \$55 (pbk); £60, \$110 (hbk)

Norman Myers

We really needed this book ten years ago when it would have illuminated an urgent but largely uninvestigated challenge of conservation biology. We have long been aware that the great bulk of the mass extinction currently under way is made up of insects, yet we have had only a meagre grasp of the details.

We have 'guesstimated' that 80% of the roughly 10 million species on the planet are insects. Yet we know so little about them that we haven't even located the main concentrations of insects (although one strong contender is the canopies of tropical forests). We know next to nothing about their natural histories or other key characteristics. And most important, we have only vague clues about their conservation status: how many species should be classified as threatened? Are species being eliminated at rates matching those for mammals and birds — that is, hundreds or even thousands of times faster than before modern humans appeared? All these questions are addressed in this compendious book.

It's true that a few taxa are well documented, notably butterflies (about 20,000 species), ants (8,000), dragonflies (6,000) and tiger beetles (25,000). But these total only some 60,000, and we cannot say how far they serve as indicator species to throw light on the rest. Fortunately, we can gain some insight by drawing on the congruence relationships of insects with plants. If we accept (gulp) that there are at least 300,000 species of plant and 8 million species of insect, that works out at one plant species

for every 27 insect species. Crude though this calculation is, it is indicative. Of course, the relationships between plants and insects are greatly varied: a few insect species rely on a single plant species, whereas many link up with dozens of plants.

Some observers may respond that if thousands of insect species are becoming extinct, so what? Do we really need all those creepy-crawlies? This point applies particularly to beetles, which must total several million species, many of them only marginally differentiated in their morphologies. Yet this apparent redundancy may serve some vital function in nature, if only as an evolutionary insurance mechanism. Still more to the point is that insects supply a host of ecosystem services that support the human enterprise. If we were eventually to lose half of all mammal and bird species, as looks entirely possible,

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

Will these *Membracis* treehoppers join countless other insect species on the road to extinction?

Taking a tip from the past

Poison Arrows: The Amazing Story of How Prozac and Anaesthetics Were Developed From Deadly Jungle Poison Darts

by Stanley Feldman

Metro: 2005. 244 pp. £14.99

John Carmody

Who would have thought that such a small book could contain so many errors? These are not the errors that a pernickety specialist dutifully finds in a popular account of a complex scientific topic. We're all prone to those, and need to keep in mind the eighteenth-century advice of Viscount Bolingbroke: "Truth lies within a little and certain compass, but error is immense." Rather, Feldman's errors in *Poison Arrows* are fundamental and imply a flawed scientific understanding of his material.

His topic is an important one but has a longer history than he seems to recognize. Curare, the paralytic arrow poison from South America, together with its derivatives and congeners, has unquestionably been a crucial element of the revolution in modern anaesthesia. But in the wider story of how nerve cells communicate with one another and with their target organs, such as muscles, glands and the heart, curare was never the central factor that Feldman — with an anaesthetist's understandable passion for these drugs — wants his readers to believe. The history of how nerve-cell communication came to be accepted as chemical in its operation, rather than electrical, has an earlier, more nebulous beginning than Feldman recognizes.

Until the late eighteenth century, conventional medical thinking assumed that good health follows from a proper balance of the four putative 'humours' of the body. Disease was a consequence of their imbalance, and we still talk of someone being in ill-humour. Then Alessandro Volta and Luigi Galvani discovered what they called 'animal electricity', and the era of what I call the 'dry brain' began. So, while pharmacologists pressed on with their potions and solutions, physiologists were, for the next century, fixated on electricity. Then Ramón y Cajal offered a revolutionary challenge. The nervous system, he asserted (on the basis of sublime microscopy), is composed of individual nerve cells (neurons) and is not a continuous network. In 1898, Charles Sherrington coined the term synapse for the intercellular links that were required to make Cajal's neurons operational, but the question remained contentious for some fifty years: is the connection electrical or chemical?

If only the physiologists had taken greater notice of their pharmacological colleagues, including Claude Bernard, who did important work with curare, as Feldman tells with some prolixity. In fact, it was an Austrian-based pharmacologist, Otto Loewi, who published

Targeting the nervous system: darts dipped in the toxin curare are used by hunters in South America.

the first convincing demonstration of chemical neurotransmission in 1921. Feldman tells little of this, favouring instead his fellow Briton Henry Dale, who shared a Nobel prize with Loewi. Feldman consistently misspells the term that Loewi chose for this transmitter, *Vagusstoff* ("Vagus nerve material"), as "vagusstuff".

That is only one of innumerable errors in this book, which betray what seems to be hasty writing and shoddy editing. No less important, Feldman rarely rises to the challenge of presenting an important scientific story. This is regrettable because many of today's therapeutic (and abused) drugs act by influencing chemical transmission between cells, either by modifying the production or disposal of the relevant chemicals or by influencing the sensitivity of the target cells to those transmitters. They symbolize the rehabilitation of the 'wet brain'.

A clear chemical understanding is necessary for both the research worker and the storyteller. Feldman's chemistry seems remarkably meagre. He refers to sodium and potassium molecules as often as he calls them ions, so it is no surprise that subtle matters, such as the importance of molecular affinities in driving reactions, are presented maladroitly. Just as bad, he seems not to recognize that when he and other anaesthetists measure muscular contractions in curarized patients or volunteers, this is an exceedingly indirect (and

potentially fallacious) means of judging how much acetylcholine — the relevant transmitter — has been released from the nerve. It is also a poor guide of the sensitivity to it of the crucial receptor molecules in the target muscle. The essential problem with this analysis is that the transmission has an enormous 'safety factor': far more molecules of acetylcholine are released than the minimum required to generate the electrical signal that will, in turn, trigger the muscle contraction. It is akin to giving someone about \$100 for a \$15 train fare and then later trying to judge the amount of money in his pocket from the evidence of whether he actually bought a ticket or not. Even worse is trying to base a plausible financial theory simply on whether or not he's on the train.

Feldman is undeterred by such details. He seems, for example, quite unaware of the fact that there are two distinct classes of receptor molecule for acetylcholine. And he overplays the importance of this transmitter in the operations of the nervous system.

The history of curare is important for several reasons, principally because this chemical, in tandem with the development of safer anaesthetic drugs, is the real reason for the success story of modern surgery. It still awaits its Edward Gibbon. ■

John Carmody is a neuroscientist and writer at the University of New South Wales in Sydney, Australia.

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

The best solution

Optimization: this beguilingly simple idea allows biologists not only to understand current adaptations, but also to predict new designs that may yet evolve.

William J. Sutherland

"If one way be better than another, that you may be sure is nature's way." Aristotle clearly stated the basic premise of optimization in biology, yet it was almost 2,000 years before the power of this idea was appreciated. The essence of optimization is to calculate the most efficient solution to a given problem, and then to test the prediction. The concept has already revolutionized some aspects of biology, but it has the potential for much wider application.

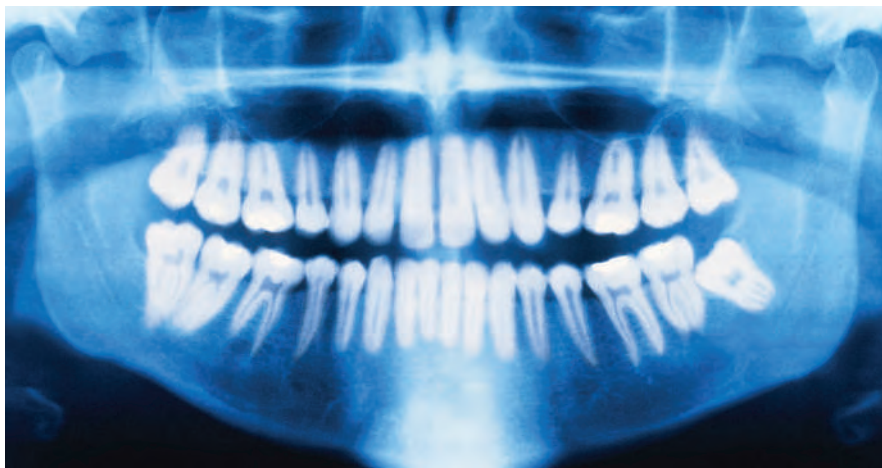
Of course, optimization has long been employed effectively in subjects other than biology. Economists have traditionally calculated the options that result in the greatest profit, and engineers routinely calculate the best design solution, such as the strongest bridge of a given weight.

Darwin's theory of natural selection provided an obvious mechanism for explaining optimization in biology: more efficiently designed individuals will leave more offspring. But it was another century before biologists calculated optimal solutions. David Lack pioneered its use in biology with his concept of the optimal clutch size — the number of eggs that would produce the greatest number of offspring.

The use of optimization has allowed biologists to move from merely describing patterns or mechanisms to being able to predict, from first principles, how organisms should be designed. Optimality models are based on three elements: the choices available; what is being optimized; and the constraints.

Physiologists have used optimization to answer a wide range of questions about animal morphology. For example, optimization has been invoked to predict the design of a bone of given weight that minimizes the risk of breaking or buckling; the speed at which it is most efficient to switch from running to walking; and the gut design that provides the highest energy gain from a given diet. The prediction of the triplet code as the most parsimonious means of coding 20 amino acids using the four bases of DNA is another successful example of this methodology.

But optimization has its critics. The most common objection centres on the mistaken belief that the aim of this method is to test whether organisms are



Revealed: optimal-design theory can be used to assess how selective forces have shaped teeth.

optimal. Actually, it is the assumptions of optimality that are tested. The failure to find support for a prediction can be used to determine whether an assumption is wrong. For example, if animals do not select the diet that maximizes energy intake, it may be because they are choosing a diet that optimizes a balance of different components, or that avoids the costs associated with obtaining larger prey. Once such possibilities have been identified, a new theory can be devised and its predictions tested. It has been argued that this process is circular but in practice it is no different from the successive predicting and testing that underlies most science.

A recent example of the insight that optimization can provide concerns the design of mammalian mouths. It is possible to predict, on the basis of efficient food fracture, how various components of the mouth, such as tooth size, should vary with body size. These predictions can then be compared with actual allometric relationships. Intriguingly, the correlation can be applied to hominid evolution: the traditional approach of predicting morphology from given constraints is reversed to consider how the constraints are likely to have resulted in the observed morphological changes.

Human mouths have become greatly reduced over the past 300,000 years, presumably as we have learnt to fragment food with tools and reduce its toughness with cooking. The predictions from optimal-design calculations are that, for a given body weight, face and incisor size should be directly proportional to the extent to which food is fragmented by

tools. Further calculations give the prediction that the reduction in molars and premolars depends on the cube root of the drop in food toughness. On the basis of these predictions, the changes caused by cooking would have to be vast to match the changes caused by tool use. As predicted, although all teeth have become reduced, the face and incisors have become proportionately smaller. This means the mouth can no longer accommodate the molars, hence the squeezed or missing third molars (wisdom teeth) of many modern humans.

A considerable strength of using optimization is that once we understand why organisms are as they are, then it should be possible to understand how they will respond to new conditions. Optimization can therefore be used to understand behaviour, and to predict population dynamics, in new environments, such as those resulting from habitat loss or a rise in sea level.

There are increasing calls for biology to be predictive. Optimization is the only approach biology has for making predictions from first principles. The wider adoption of these ideas right across biology should reap ample rewards.

William J. Sutherland is at the Centre for Ecology, Evolution and Conservation, School of Biological Sciences, the University of East Anglia, Norwich NR4 7TJ, UK.

FURTHER READING

Alexander, R. M. *Optima for Animals* (Princeton Univ. Press, Princeton, 1996).
Lucas, P. W. *Dental Functional Morphology* (Cambridge Univ. Press, Cambridge, 2004).
Sutherland, W. J. *From Individual Behaviour to Population Ecology* (Oxford Univ. Press, Oxford, 1996).

STONE/GETTY IMAGES

CONCEPTS

NEWS & VIEWS



S. CHEN/CORBIS

HUMAN BEHAVIOUR

Brain trust

Antonio Damasio

As is the case with other social interactions, financial transactions depend on trust. That fact is behind ingenious experiments that explore the neurobiological underpinnings of human behaviour.

Michael Kosfeld and his colleagues got students in Zurich to play a serious game. The game involved real monetary exchanges between two people playing the anonymous roles of 'investor' and 'trustee'; beforehand, each subject had received either the neuropeptide oxytocin or an inert placebo, via nasal spray.

As a group, the investors who received oxytocin exhibited more trust in the anonymous trustee than did the investors who received the placebo. Because intranasally administered oxytocin crosses the blood-brain barrier into the central nervous system, Kosfeld *et al.* (page 673 of this issue)¹ conclude that the central action of oxytocin increases trusting behaviour; and because the oxytocin spray did not change the behaviour of the trustees, it seems that oxytocin only increases trust, not the reliability of the trustee. This is a remarkable finding, and to explain its significance we must first say a word about trust and about oxytocin itself.

Given the polarities of reward and punishment that pervade biology at various levels, trust is essential for the normal operation of human societies. Remove trust and you compromise love, friendship, trade and leadership. Little is known about the neurobiology of trust, although the phenomenon is beginning to attract attention².

As for oxytocin, it is a small peptide, consisting of nine amino acids, that is produced

mostly in the hypothalamus, the brain's master controller of biological regulation, including emotion. Oxytocin acts both on certain targets of the body (it is best known for inducing labour and lactation) and on brain regions whose function is associated with emotional and social behaviours (the amygdala and nucleus accumbens, for example) — that is, it works both as a hormone and as a neuromodulator, a kind of neurotransmitter. In animals, oxytocin contributes to social attachments, including male and female bonding after mating, mother and infant bonding after childbirth, and assorted sexual behaviours^{3,4}. Besides triggering complex and specific action-programmes, oxytocin may well work part of its charm by selectively lowering the natural resistance that animals have to the proximity of others, thus facilitating what is known as 'approach behaviour'.

Given this background, Kosfeld *et al.*¹ hypothesized, reasonably and perceptively, that oxytocin might be involved in trusting behaviour in humans. After all, trust and approach behaviour are indelibly linked. We commonly describe the child who approaches others with ease as 'trusting', and we use comparable descriptions for animals in similar situations. Kosfeld and colleagues' finding supports their hypothesis and opens the way to a richer understanding of perhaps the most complex

tier of human social interactions. I once likened⁵ oxytocin to a love potion, the magic elixir that makes Tristan fall for Isolde: add trust to the mix, for there is no love without trust.

Kosfeld *et al.* provide an engaging discussion of the possible mechanisms behind their finding. They reject the possibility that oxytocin has a nonspecific positive effect on social behaviour, because of its different influence on investors and trustees. Approach and trust possibly dominate the behaviour of investors, and that is where oxytocin works, whereas trustee behaviour is dominated by a principle of reciprocity, for which oxytocin seems irrelevant. Kosfeld *et al.* also reject the possibility that oxytocin merely reduces the sensitivity to risk, because in a control experiment in which the investors knew the trustee was a computer, they did not take any extra risks. The authors finally settle for an attractive pair of factors: that oxytocin overcomes the aversion to betrayal (which applies only to the investors), and that this is combined with the effects of reward that result from enhanced approach behaviour.

The significance of the study lies in what it can tell us about non-experimental circumstances, when the equivalent of an investor is not sniffing oxytocin. What might be happening then? First, perceiving certain social configurations probably leads to oxytocin release

in selected brain regions — that is, the cognitive appraisal of a situation, based on an individual's genetic make-up and past experience, triggers a chain of neural events that includes (but is not limited to) the release of oxytocin. Second, oxytocin modulates the activity of cognitive neural networks, resulting in enhanced trusting behaviour. Whether this result is achieved via a mostly unconscious bias (by altering the competition among ensembles of neurons that represent varied choice options), or a conscious deliberative process, remains to be established — although the evidence seems to favour the former possibility in the current experiment. However, the input, along the cognitive chain, of neural events arising in brain areas associated with social and emotional responses is a requisite part of the explanation. The finding points to the crucial involvement of emotional phenomena in the processes leading from cognition to behaviour.

The authors' results open up possibilities for investigating conditions in which trust is either diminished, as in autism, or augmented. For example, patients with bilateral damage to the amygdala approach strangers with unusual ease, and fail to recognize untrustworthy individuals whom normal people would resolutely avoid⁶. In this case, damage to the amygdala may prevent the detection of the potential threat evoked by certain stimuli. And children with Williams syndrome, a rare genetic disorder, approach strangers fearlessly and indiscriminately⁷. Might their high level of trust be due to excessive oxytocin release?

Some may worry about the prospect that political operators will generously spray the crowd with oxytocin at rallies of their candidates. The scenario may be rather too close to reality for comfort, but those with such fears should note that current marketing techniques — for political and other products — may well exert their effects through the natural release of molecules such as oxytocin in response to well-crafted stimuli. Civic alarm at the prospect of such abuses should have started long before this study, and the authors cannot be blamed for raising it. Whatever the beneficial biomedical applications, or the abuses, may turn out to be, Kosfeld *et al.* have made a valuable contribution to our understanding of the role of neuromodulators in human behaviour that involves choice. ■

Antonio Damasio is in the Department of Neurology, University of Iowa College of Medicine, 200 Hawkins Drive, Iowa City, Iowa 52242, USA.
e-mail: antonio-damasio@uiowa.edu

- Kosfeld, M., Heinrichs, M., Zak, P. J., Fischbacher, U. & Fehr, E. *Nature* **435**, 673–676 (2005).
- King-Casas, B. *et al.* *Science* **308**, 78–83 (2005).
- Carter, C. S. *Psychoneuroendocrinology* **23**, 779–818 (1998).
- Insel, T. R. & Shapiro, L. E. *Proc. Natl Acad. Sci. USA* **89**, 5981–5985 (1992).
- Damasio, A. R. *Descartes' Error: Emotion, Reason, and the Human Brain* (Penguin, New York, 1994).
- Adolphs, R. & Damasio, A. R. *Nature* **393**, 470–474 (1998).
- Doyle, T. F. *et al.* *Am. J. Med. Genet.* **124A**, 263–273 (2004).

COSMOLOGY

Digitizing the Universe

Nickolay Y. Gnedin

For years, cosmologists have been racing each other to develop ever more sophisticated and realistic models of the evolution of the Universe. The competition has just become considerably stiffer.

Since the first 'analogue' simulation by Erik Holmberg¹, who used the inverse-square law of light to mimic gravity, numerical cosmology has made remarkable progress: abstract particles and digital supercomputers have now replaced light bulbs and photometers as tools for measuring gravitational forces. But the ultimate dream of every cosmologist — to create a realistic model of the whole Universe inside a computer — remains elusive. The Universe is just too complicated and too large for even the fastest supercomputers.

So computational astrophysicists have to invent clever shortcuts. They hide physics that is too poorly known, or too complex to be modelled from first principles, in phenomenological (often called 'semi-analytical') models. On page 629 of this issue, Springel *et al.*² describe the best example of this approach to date, which they appropriately named the Millennium Run (if we forgive the overrun of a few years, it was indeed the largest and most realistic simulation of the last millennium).

Springel and colleagues² — an international collaboration of computational astrophysicists known as the Virgo Consortium — adopted arguably the best approach available to model the formation and evolution of galaxies and quasars in a representative volume of the Universe. They used an ultra-high-resolution simulation to follow in unprecedented detail the evolution of dark matter — invisible material that is the dominant source of gravity in the Universe. Ironically, although the nature of dark matter remains unknown, its clustering properties (the ways it clumps together into 'haloes') are rather well understood, thanks to a substantial amount of observational data gathered in the past decade, and to ever more accurate numerical simulations. At present, cosmologists can simulate dark matter, which we can't see, better than galaxies and gas, which we can.

The Virgo Consortium therefore combined the best dark-matter simulation to date with a semi-analytical, but physically motivated, model of the formation of stars and super-massive black holes. These latter, despite being dubbed 'black', are actually the brightest objects in the Universe — 'quasars'. Sitting at the centres of galaxies, quasars accrete large quantities of very hot gas, which emits enormous amounts of radiation just before being swallowed for ever by a black hole. The consortium's end-product — the Millennium Run — gives us the most detailed and accurate

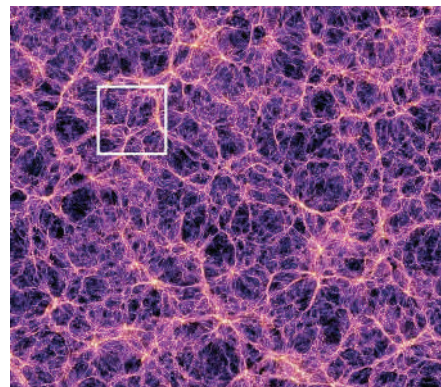


Figure 1 | Simulated cosmos. The density distribution of matter in a slice of the computational volume of the Millennium Run model, showing large clusters with densities 1,000 times the mean density of the Universe (yellow); a 'cosmic web' of filamentary structures 10 to 100 times denser than the mean (purple); and the mostly empty regions (black), often called voids, which contain less than 10% of the mean density of the Universe. The white square shows the size of the computational volume for a full hydrodynamic simulation that would use up the same computational resources as the Millennium Run. (Figure courtesy of Volker Springel.)

theoretical prediction so far of the properties of galaxies and quasars, from the dawn of cosmic time to the present.

An alternative approach (of which I am a devout adherent) to this kind of semi-analytical modelling is to use a hydrodynamic simulation code, which follows the complex swirl of cosmic gas from the largest scales down to the tiny and ultra-dense molecular clouds where stars are born. Unfortunately, however, the vastly greater numerical complexity of such calculations means that the direct hydrodynamic approach cannot, at present, be used to model a representative piece of the Universe for the whole duration of cosmic evolution. The extent to which hydrodynamic simulations lag behind the Millennium Run in this respect is illustrated in Figure 1.

What could the Millennium Run be used for? Springel *et al.*² have only scratched the surface, but the agreement between the simulation and observational results is already stunning. To take just one example, the Sloan Digital Sky Survey³ (SDSS), a project that aims to map in detail a quarter of the sky, continues to discover ever more distant quasars⁴. The current record holder is a truly remarkable object. Taking account of the time its light has

taken to reach us, the quasar we see has had only about 850 million years to grow since the Big Bang, and yet its luminosity is a dazzling 10^{13} times that of the Sun. The mere existence of such an object, some argue, spells serious trouble for the currently favoured theory of cosmic structure formation, the 'cold dark matter' (CDM) model.

This model is known as such because the dark matter is assumed to comprise massive, slowly moving — hence 'cold' — elementary particles. In this model, cosmic structure forms hierarchically, with small objects merging to form larger ones. Some researchers naturally question whether 850 million years is long enough to form an object such as the most distant SDSS quasar. A quasar that can outshine ten trillion Suns would have to harbour a black hole of almost a billion solar masses, and forming such a monster in only 850 million years is a daunting task. But one of the most remarkable conclusions of the Millennium Run is that such objects form naturally in the CDM model in numbers consistent with the SDSS data, despite the fact that many smaller objects must merge in a relatively small region of space.

The Millennium Run is not, of course, limited to modelling distant quasars. As a comprehensive and realistic model of the Universe, it can be used to answer an array of questions about the properties of individual galaxies and how they clump together today and in the earlier stages of cosmic evolution. As Springel *et al.*² show, the Millennium Run is an adequate theoretical tool for making thorough comparisons with the best observational data available, such as those provided by the SDSS. Although a large amount of work still lies ahead for the Virgo Consortium, it is rewarding to think that, armed with modern computing power, we theorists can keep up with the best observational advances — even though, in pure manpower, the observers outnumber us three to one.

Nickolay Y. Gnedin is at the Center for Astrophysics and Space Astronomy (CASA), University of Colorado, Boulder, Colorado 80309, USA.
e-mail: gnedin@casa.colorado.edu

1. Holmberg, E. in *Annals of the Observatory of Lund* **6** (1937).
2. Springel, V. *et al.* *Nature* **435**, 629–636 (2005).
3. www.sdss.org
4. Fan, X. *et al.* *Astron. J.* **125**, 1649–1659 (2003).

STEM CELLS

The road not taken

Hanno Hock and Stuart H. Orkin

Developmental 'road maps' chart the steps from simple cells to mature, specialized cells. A newly discovered variety of blood-cell progenitor doesn't fit into the accepted blood map, but should that map be redrawn?

The diverse specialized cell types of the body are all derived from comparatively rare, immature cells known as stem cells. This process is almost always represented by hierarchy diagrams, in which cells are depicted making orderly decisions, progressively restricting their developmental options with each step. Because such diagrams organize complex biological information, they tend to be readily accepted and have come to dominate our ideas. Adolfsson *et al.*¹, writing in *Cell*, challenge the commonly accepted map for blood-lineage development and remind us that our knowledge of even this well-studied system is in fact incomplete and evolving.

Rare stem cells that reside in the bone marrow (the haematopoietic stem cells, or HSCs) are the source of the various blood-cell types. The classic experiments of Weissman and colleagues^{2,3} showed that HSCs give rise to progenitor populations that can be isolated prospectively by using antibodies against cell-surface proteins. Analysis of different progenitor populations suggested a simple hierarchy with progressively restricted developmental potential (Fig. 1a, overleaf). The result is an

alphabet soup of progenitor cells — the CLPs, CMPs, GMPs and MEPs — that is now engrained in the jargon of stem-cell biology, haematology, immunology and leukaemia research.

The common lymphoid progenitor (CLP) develops into T and B cells, the white blood cells that mediate acquired immune responses against pathogens. The CLP is distinct from the common myeloid progenitor (CMP), the source of all the 'myeloid' lineages — red blood cells (also called erythrocytes, the oxygen transporters), megakaryocytes (off which platelets bud), and white blood cells called monocytes and granulocytes (which mediate inflammation and carry out innate immune responses). Myeloid cells mature from CMPs through two intermediates: the megakaryocyte/erythrocyte progenitor (MEP) and the granulocyte/monocyte progenitor (GMP).

The frequencies, developmental potentials and gene expression profiles of the blood progenitor cells have been reported in countless papers dealing with normal and abnormal situations in mouse models and human patient samples. As a result one can now, for

example, superimpose on the simple haematopoietic map the expression pattern of crucial nuclear regulatory proteins to derive a superficial view of how developmental decisions are executed as the HSCs mature into the individual blood lineages⁴.

Adolfsson and colleagues' results¹ cast doubt on the widely held belief that the first bifurcation from HSCs to the various lineages distinguishes the lymphoid and myeloid pathways from each other. They have identified an adult bone-marrow population (termed LMPP, for lymphoid primed multipotent progenitor) that has the potential to become both lymphoid and granulocyte/monocyte lineage cells but curiously lacks the capacity to generate red cells or megakaryocytes (Fig. 1b).

HSCs and LMPPs have a similar profile of stem-cell-specific surface proteins (expressing c-Kit and Sca-1, but no markers for differentiated blood lineages). However, they differ in their expression of the surface receptor kinase Flt3. The Flt3 protein is required for CLP development, and its presence also signifies the loss of the capacity for long-term self-renewal, one of the hallmarks of HSCs^{5–8}. Adolfsson *et al.* show a correlation between the upregulation of Flt3 and the loss of erythroid and megakaryocytic potential.

These findings imply that there is no single, obligatory path from HSCs to granulocytes or monocytes. The authors propose two potential models, one very provocative, the other more easily reconciled with current dogma. The first posits that MEPs arise from the HSC (or an immediate descendant), whereas GMPs arise exclusively from the LMPP (Fig. 1c). In stark contrast to current views, this model argues against the existence of true CMPs (cells that can give rise to all myeloid but no lymphoid progeny). In its most extreme form, this model is not well supported by available data. Akashi *et al.*³ previously showed that 62% of purified CMPs give rise to colonies containing both granulocyte/monocyte and megakaryocyte/erythroid lineage cells. In principle, Adolfsson *et al.*¹ confirm the existence of functional CMPs even though, perhaps because of subtle differences in technique, they find that CMPs fulfil their predicted potential at a much lower frequency (8%). The strength of this model is that it highlights the unexpected nature of the findings and defines the issues that need to be resolved.

The second model, which seeks to consolidate the new data with previously established findings, proposes that LMPPs coexist with CMPs, thereby providing two alternative sources of GMPs (Fig. 1d). However, it remains to be formally demonstrated that LMPPs do indeed give rise to GMPs rather than representing an entirely different maturation pathway. Another caveat is that the model seems to intimate that the proportion of GMPs arising through the alternative pathway equals that developed through the conventional route, but whether this is actually the case is unclear.

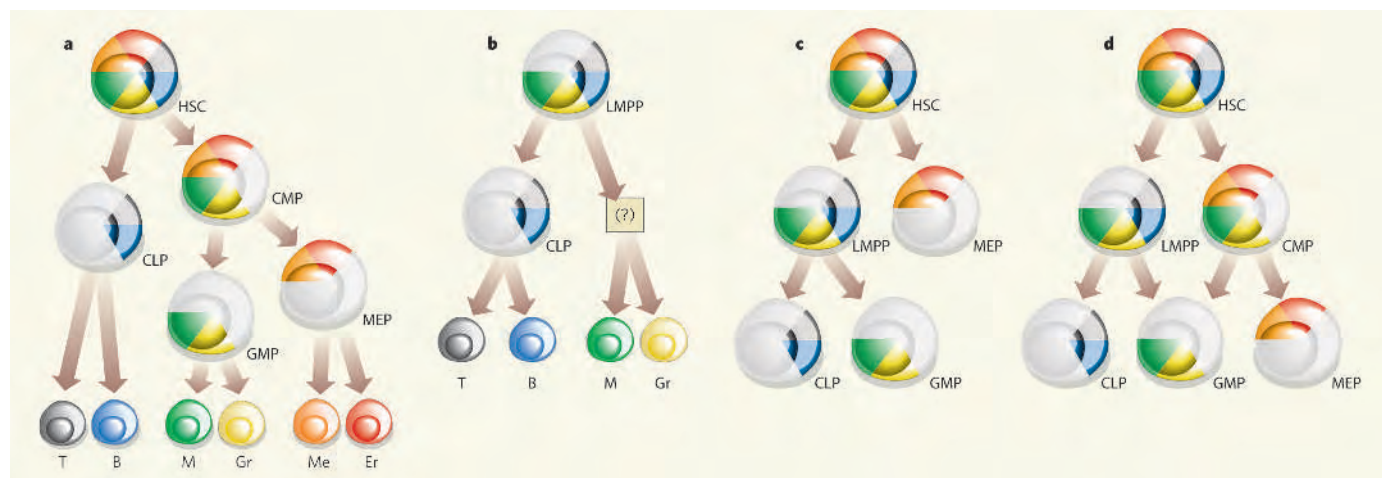


Figure 1 | Blood lines — cell-fate decisions of haematopoietic stem and progenitor cells. **a**, Analysis of progenitor populations initially suggested that the first developmental step leads to committed common lymphoid and myeloid progenitors, CLP and CMP, respectively. **b**, Adolfsson *et al.*¹ discovered a progenitor population (LMPP, lymphoid primed multipotential progenitors) that does not fit the established scheme (**a**) because it has lost

erythroid and megakaryocytic potential but can still give rise to granulocytes and macrophages, as well as T and B cells. Colours in the various progenitor cells symbolize their developmental potential after purification to single cells. **c**, **d**, Two alternative models proposed by Adolfsson *et al.* for the position of the LMPP in the haematopoietic map. Er, erythrocyte; Me, megakaryocyte; M, monocyte; Gr, granulocyte; T, T cell; B, B cell.

Is it time to discard the widely used haematopoietic lineage map? Not really, but it is important to remain aware that it organizes complex data at the risk of simplification. Progenitors that fail to fit snugly into the simple scheme have been reported previously. For instance, fetal liver progenitors can be 'bipotent' for macrophages and B cells, or 'tripotent' for macrophages, B and T cells^{9,10}. In addition, the 'lineage infidelity' that is evident in some malignancies may not merely reflect aberrant development programmed by cancer-promoting genes, but may rather provide a glimpse into progenitors that don't fit comfortably into simple diagrams.

Before full commitment to a particular lineage, the nuclear regulators of diverse lineages are expressed in immature cells, a phenomenon known as lineage priming¹¹. Static hierarchy diagrams are inadequate to describe the potential instability and plasticity of immature progenitors with regard to their ultimate lineage affiliation. As haematopoietic progenitors are separated further into additional subpopulations, we are likely to be faced with other examples that violate conventional logic. Previous experience teaches that lessons learned from haematopoiesis also apply to other organ systems, so we should be prepared to entertain further detours along the highway of development. ■

Hanno Hock and Stuart H. Orkin are at the Children's Hospital and Dana-Farber Cancer Institute, Harvard Medical School and Howard Hughes Medical Institute, Boston, Massachusetts 02115, USA.
e-mail: stuart_orkin@dfci.harvard.edu

5. Mackarehntschian, K. *et al. Immunity* **3**, 147–161 (1995).
6. Adolfsson, J. *et al. Immunity* **15**, 659–669 (2001).
7. Christensen, J. L. & Weissman, I. L. *Proc. Natl Acad. Sci. USA* **98**, 14541–14546 (2001).
8. Sitnicka, E. *et al. Immunity* **17**, 463–472 (2002).

9. Cumano, A., Paige, C. J., Iscove, N. N. & Brady, G. *Nature* **356**, 612–615 (1992).
10. Lu, M., Kawamoto, H., Katsube, Y., Ikawa, T. & Katsura, Y. *J. Immunol.* **169**, 3519–3525 (2002).
11. Orkin, S. H. *Immunity* **19**, 633–634 (2003).

MOLECULAR ELECTRONICS

Charged with manipulation

Mark Ratner

The ability to control charge transport through individual molecules sandwiched between electrodes could lead to further miniaturization of electronics. A better understanding of how such junctions work is crucial.

In its simplest form, a molecular junction consists of a single organic molecule sandwiched between two much larger electrodes. More efficient and more precise control of current flow through such elements, and a detailed understanding of the factors that influence that flow, are essential to take electronics from the microscale of conventional, silicon-based technology down to the nanoscale. On page 658 of this issue¹, Wolkow and colleagues demonstrate how to control the onset of conduction through a molecular junction by using the charge state of an atom on the surface of a silicon electrode.

The difficulty with structures such as molecular junctions is that, although the molecule consists of a series of discrete states in a small, finite entity, the electrodes contain a very dense set of states in a macroscopic structure. Understanding how the electrostatic environment of the molecule modifies the transport process — in other words, the influence of the electrode — is, in many ways, the most vexing problem in dealing with such junctions.

In the more typical metal–molecule–metal

type of molecular junction^{2–6}, extra complexity results from the geometric disorder inherent in the usual sorts of metal–molecule coordination bonding. Typically, these bonds are between a coinage metal (copper, silver or gold) and a thiol group, or between palladium and a cyano group⁷. Geometric changes can strongly affect charge transport through these junctions and give rise to the random switching phenomena often seen^{8–11} in such structures.

One way to avoid this geometric uncertainty is to exploit transport in junctions built not on a metal but on a semiconductor (particularly silicon), where the molecule–electrode interface can be provided by two atoms 'sharing' the electrons in a covalent bond. This bond can be created in several ways^{12–16}, but perhaps most directly by linking a free radical at the end of the molecule — in this case, a carbon atom with an unpaired electron — to a 'dangling bond' on the surface of the silicon electrode. This dangling bond can arise through the removal of a hydrogen atom originally attached to each silicon atom at the surface of

1. Adolfsson, J. *et al. Cell* **121**, 295–306 (2005).
2. Kondo, M., Weissman, I. L. & Akashi, K. *Cell* **91**, 661–672 (1997).
3. Akashi, K., Traver, D., Miyamoto, T. & Weissman, I. L. *Nature* **404**, 193–197 (2000).
4. Orkin, S. H. *Nature Rev. Genet.* **1**, 57–64 (2000).



50 YEARS AGO

"Origin and Age of Meteorites." In his recent article under this title, Prof. H. C. Urey says: "While the excellent studies of Paneth and his co-workers are most interesting because of the facts which they present about the formation of meteorites and the effects of cosmic rays, I believe that it is very difficult to decide what event or events were recorded by the helium, uranium and thorium abundances."

We should like to put on record that we agree with Prof. Urey that there are difficulties in the interpretation of our results; but that we do not think that his statements prove that our arguments, leading to an age of no more than a few hundred million years, are incorrect. We are not convinced that his interesting thermodynamic reasoning can be applied to the calculation of the maximum quantities of uranium and thorium in iron meteorites, since it is by no means certain that their distribution was achieved under equilibrium conditions.

F. A. Paneth *et al.*

From *Nature* 4 June 1955.

100 YEARS AGO

The thirty-fifth of the privately printed opuscula... of the Sette of Odd Volumes is entitled "The Early History of the Royal Society." The author of this brochure is Mr. Henry B. Wheatley... who has succeeded in writing a very interesting account of the early years of our national association of men of science. Mr. Wheatley shows that Charles II — "Founder, Patron, and one of the Royal Society of London for improving Natural Knowledge" — took a genuine interest in the advancement of the society. "True he did not give any money, but then money was never very plentiful with His Majesty."... Objections were on one occasion made to Charles II that a member recommended by him for election was a shopkeeper. By way of reply the King "gave this particular charge to his Society, that if they found any more such tradesmen they would be sure to admit them all, without any more ado."

From *Nature* 1 June 1905.

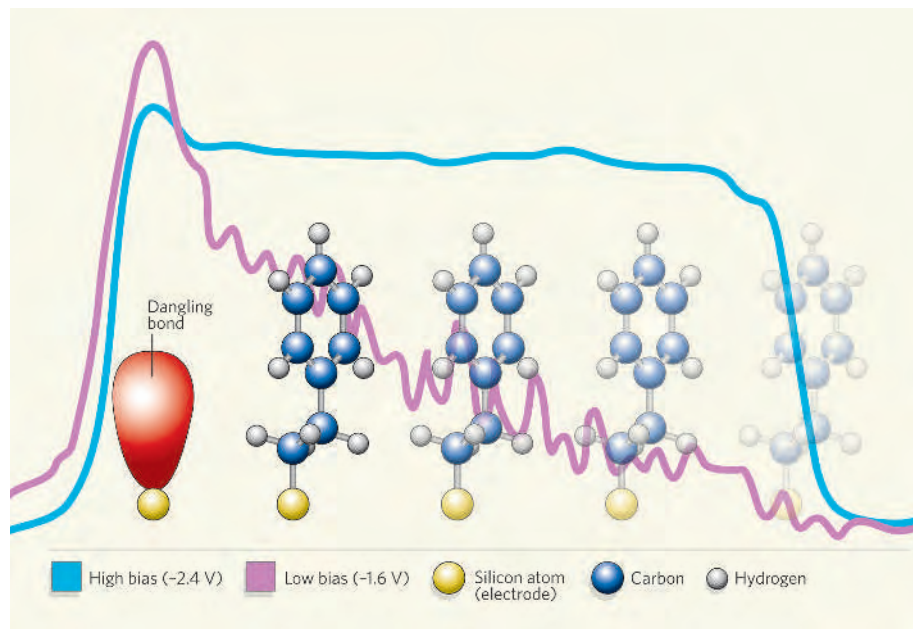


Figure 1 | Dangling potential. The polymerization of a molecule (here an organic styrene-derived molecule, not to scale) on a silicon substrate stops abruptly at a dangling-bond site. The blue and purple lines indicate the height of the molecules, as seen by a scanning tunnelling microscope (STM) — a measure of the charge transport across the molecules. At higher bias (blue line), all molecules are 'turned on', and appear bright in the STM picture. At lower bias (purple line), all molecules should appear dark. Wolkow *et al.*¹, however, discover that the electrostatic potential of the negatively charged dangling bond causes the nearest molecules to remain bright. This suggests that such structures could be used to manipulate charge transport through molecular junctions. (Figure adapted from Fig. 1 of ref. 1.)

the passivated electrode, leaving behind an unpaired electron that hangs free.

Wolkow and colleagues¹ take advantage of two remarkable properties of silicon surfaces to characterize how changes in the charge state of a silicon surface atom influence the effective field over the molecular junction. First, the 'polymerization' of the molecules — the process by which they attach themselves covalently along 'dimer rows' on the silicon surface to form a line — ends abruptly at a dangling-bond site (Fig. 1). Second, the dangling-bond site itself can become more or less charged depending on the doping level of the silicon (that is, the deficiency or excess of electrons that is induced by adding an 'impurity element' with an intrinsic number of valence electrons different from silicon's four).

Wolkow and colleagues¹ used a scanning-probe microscope to examine the effective charges along the length of the polymers originating from a given dangling-bond site. They found that when the dangling-bond site is charged, a 'slope' structure is seen that is absent when the site is uncharged (Fig. 1): molecules attached to the silicon electrode close to a charged dangling bond appear to stand out farther from the electrode than those at a greater distance from the dangling-bond site. Wolkow and colleagues interpret this as the effect of a local electrostatic charge on charge transport through the polymeric wire. This interpretation is supported by calculations showing changes in molecular orbital states caused by the charged dangling-bond site.

This result constitutes direct evidence that localized charges profoundly affect charge transport in single-molecule structures on silicon surfaces at room temperature.

The silicon-organic interface investigated by Wolkow and co-workers has many advantages over the more conventional metal-molecule interface (at least as the latter is usually implemented). One is the tight geometric constraint provided by a single covalent bond. Another is the bandgap character of the silicon (requiring an electron to acquire additional energy to contribute to conduction), which can be changed by doping to adjust the conductivity. This bandgap can result in useful phenomena, such as the negative differential resistance^{12,17} (equivalent to an increase in voltage leading to a decrease in current) observed in transport through molecules bonded to dangling-bond sites.

More generally, changing the electrostatic potential on a molecule will change its conduction characteristics^{3,5,6,17,18}. In ordinary transistor geometries, this electrostatic control is provided by one of three electrodes, the 'gate', which regulates the amount of current that can flow from 'source' to 'drain' through the main channel of the transistor. In most single-molecule transport measurements, for example using 'break junctions'^{9,19,20}, such gating can also be attained, but it requires very large voltages. In effect, this is because the molecular entities are a long way from the gate compared with the source-drain distance. This new work suggests that more effective

electrostatic control could perhaps be obtained using localized charges.

The article by Wolkow and colleagues is one of a series of contributions^{1,12,13,16,17} that underline the potential of molecular transport junctions using silicon electrodes, and explore the different control mechanisms and mechanistic transport behaviours that can be observed in them. Using a semiconductor electrode substantially widens the perspective on single-molecule transport structures, the basic unit of molecular electronics.

Mark Ratner is in the Department of Chemistry, Northwestern University, 2145 Sheridan Road, Evanston, Illinois 60208-3113, USA.
e-mail: ratner@chem.northwestern.edu

1. Piva, P. G. *et al.* *Nature* **435**, 658–661 (2005).
2. Bowler, D. R. *J. Phys. Condens. Matter* **16**, R721–R754 (2004).
3. Nitzan, A. & Ratner, M. A. *Science* **300**, 1384–1389 (2003).
4. Reed, M. A. & Takhee, L. (eds) *Molecular Nanoelectronics*

- (American Scientific, Stevenson Ranch, CA, 2003).
5. Nitzan, A. *Annu. Rev. Phys. Chem.* **52**, 681–750 (2001).
6. Datta, S. *Quantum Transport: Atom to Transistor* (Cambridge Univ. Press, 2005).
7. Tour, J. M. *Molecular Electronics: Commercial Insights, Chemistry, Devices, Architecture and Programming* (World Scientific, River Edge, NJ, 2003).
8. Xiao, X. Y., Xu, B. Q. & Tao, N. J. *Nanoletters* **4**, 267–271 (2004).
9. Mayor, M. & Weber, H. B. *Angew. Chem. Int. Edn Engl.* **43**, 2882–2884 (2004).
10. Basch, H., Cohen, R. & Ratner, M. A. *Nanoletters* (in the press).
11. Lewis, P. A. *et al.* *J. Am. Chem. Soc.* **126**, 12214–12215 (2004).
12. Guisinger, N. P. *et al.* *Nanoletters* **4**, 55–59 (2004).
13. Tong, X., DiLabio, G. A. & Wolkow, R. A. *Nanoletters* **4**, 979–983 (2004).
14. Hamers, R. J. *et al.* *Acc. Chem. Res.* **33**, 617–624 (2000).
15. Filler, M. A. & Bent, S. F. *Prog. Surf. Sci.* **73**, 1–56 (2003).
16. Hersam, M. C., Guisinger, N. P. & Lyding, J. W. *Nanotechnology* **11**, 70–76 (2000).
17. Rakshit, T. *et al.* *Nanoletters* **4**, 1803–1807 (2004).
18. Xue, Y. Q. & Ratner, M. A. *Int. J. Quant. Chem.* **102**, 911–924 (2005).
19. Liang, W. J. *et al.* *Nature* **417**, 725–729 (2002).
20. Park, J. *et al.* *Nature* **417**, 722–725 (2002).

extending previous studies of this strategy^{2,3}, are a considerable step towards a successful genetic-engineering approach to treating human disease.

Most gene-transfer strategies are in essence gene-addition therapies: a normal copy of a gene is transferred into the cell type of interest, where it coexists with the mutant copy of the same gene. But the ultimate aim is to correct the mutation, and thus enable the production of a normal protein under the control of the cell's intrinsic regulatory signals. So far this goal has remained out of reach for human therapeutics because it is difficult to correct genes at a high enough frequency to be of clinical benefit.

Site-specific correction of a mutation can be divided into three distinct problems. First, the site to be corrected must be detected among the three billion base pairs of the genome. Second, the specific base change must be installed. And third, the 'detecting and correcting' reagents must be efficiently delivered to the relevant cell type in the body (for example, lung cells in patients with cystic fibrosis). To solve the first problem, Urnov *et al.*¹ have exploited zinc-finger nucleases (ZFNs). These synthetic proteins were first described by Chandrasegaran and colleagues⁴. Developed as reagents to cleave DNA at a specific site, they are composed of a DNA-binding domain plus a DNA-cleaving domain.

The DNA-binding domain of a ZFN comprises a string of 'zinc-finger motifs', each a stretch of around 30 amino acids, stabilized by a zinc ion, that binds to a particular three-base DNA sequence. The zinc-finger amino-acid sequence varies according to the DNA sequence to which it binds. There are 900 zinc-finger-bearing proteins encoded in the human genome⁵, affording an extensive working catalogue for the assembly of ZFNs that bind

GENE THERAPY

The moving finger

Katherine A. High

DNA-cleaving enzymes trigger a repair process that can now be harnessed to correct mutations in the human genome *in vitro*. This represents another step towards gene-correction strategies for treating human disease.

Genetic engineering and gene-transfer technologies have produced a wealth of new ideas about how to treat genetic diseases. Many of these have been applied quite successfully in cultured cells and even in mouse models of human disease, but have proven remarkably difficult to translate into clinical practice. On

page 646 of this issue, however, Urnov and colleagues¹ report a new strategy that combines two biological processes that are highly evolutionarily conserved to correct a genetic mutation in human cells. Their success rate is theoretically high enough to result in clinical benefit. Their findings, building on and

ANALYTICAL CHEMISTRY

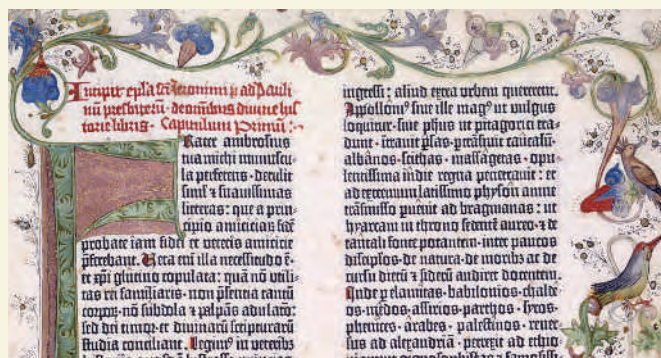
The Renaissance palette

The invention of printing using movable metal type — at least in Europe — is attributed to Johann Gutenberg of Mainz, in present-day Germany. His discovery profoundly influenced Western culture by allowing, in theory, the 'mass' reproduction of religious and secular texts. In practice, however, the printed copies were distributed to workshops throughout Europe to be lavishly embellished, or 'illuminated', by hand with headlines and capital letters.

Of the estimated 180 copies of the Bible that Gutenberg produced in three years up to 1455, substantial fragments of 48 survive. Tracey D. Chaplin *et al.* (*Anal. Chem.* doi:10.1021/ac050346y) have used

Raman spectroscopy — shining laser light onto a sample and analysing the spectrum of the scattered light — to study the chemical composition of the pigments used in the illuminations of seven copies of the Bible from libraries in France, Germany and Britain. Although the style of illumination varies widely between the Bibles, the chemical composition of the pigments used is remarkably similar, pointing to a shared knowledge of dyestuffs across fifteenth-century Europe.

Nine pigments are present, for example, in the King George III Bible (pictured) held in the British Library (pictured) held in the British Library in London. The chemical composition of seven could be identified

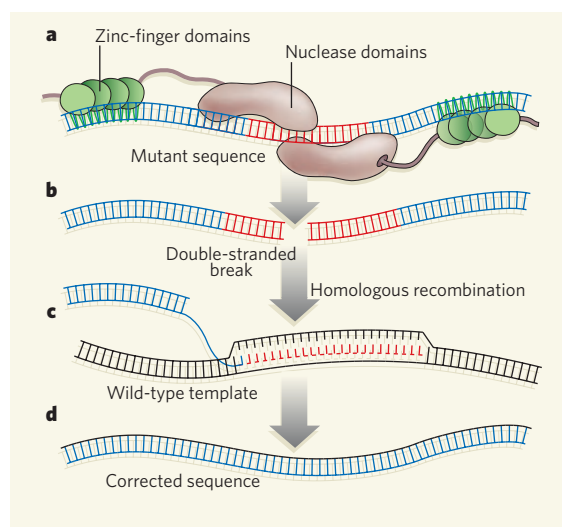


conclusively: red (cinnabar or mercury sulphide), blue (azurite, a basic copper carbonate), olive green (malachite, another basic copper carbonate), dark green (verdigris, an organo-copper compound), yellow (lead-tin oxide), black (carbon) and white (calcium carbonate). Copies of the Bible from two German libraries, printed on vellum, also contain the

expensive blue pigment lazurite (lapis lazuli) — perhaps indicating commissions from wealthy families.

The authors point out that, alongside the historical interest of their findings, an accurate knowledge of pigment make-up is crucial in helping to preserve and restore such cultural treasures.

Richard Webb

**Figure 1 | Gene correction.**

Patients with severe combined immunodeficiency bear a particular mutation that influences the function of their T cells. Urnov *et al.*¹ used the strategy shown here to correct the fault in a significant percentage of cultured cells carrying the same mutation. **a**, Zinc-finger nucleases are designed so that their zinc-finger domains will recognize specific sites near the mutation. **b**, The nuclease domains introduce a double-strand break into the DNA. **c**, A wild-type (unmutated) sequence is introduced into the cells and used as the template for a cellular repair process, homologous recombination. **d**, The mutant sequence is corrected.

to different sequences. And this catalogue has been expanded using a technique called phage display⁶.

The DNA-cleaving domain of ZFNs is derived from the enzyme FokI, and is not itself specific. As this domain must dimerize to achieve efficient DNA cleavage, the strategy requires two ZFNs to bind at or near the site to be corrected. By linking four zinc fingers in tandem for each of the two ZFNs, a combined recognition site of 24 base pairs is attained, specifying a unique address within the genome (Fig. 1). So the first key reagent, allowing detection of the site to be corrected, is already available.

This dimerized enzyme introduces a double-strand break into DNA at or near the site of a mutation. The cell's powerful repair mechanisms are then engaged. One of these, 'homologous recombination', repairs the damage by using similar sequences within the cell as a template, in a process that involves switching DNA strands between stretches of base-paired DNA. The similar sequence would usually be found within the genome, but it could also be introduced into the cell. This solves the second problem — effecting a base change.

The most remarkable feature of this complex process is the frequency at which it occurs, even in the absence of any selection for corrected cells. Urnov *et al.*¹ show correction of 15–20% of mutated chromosomes in a cell line (that is, in serially cultured cells), allowing the efficient introduction of two engineered ZFNs and the normal template to effect the change. They even report around 5% correction in human T cells taken from a subject, where introduction efficiencies are usually considerably lower. This contrasts starkly with the rate of homologous recombination when no double-strand breaks are introduced, which is closer to 0.001%.

What implications does this work have for treating human disease? The DNA sequence that Urnov *et al.* altered is one that is mutated in one form of severe combined

immunodeficiency (SCID) — a disease that, without bone-marrow transplantation or gene therapy, is fatal early in childhood. Several years ago, Cavazzana-Calvo *et al.*⁷ reported successful gene-transfer studies of SCID. They used a retrovirus to add a normal gene to target cells (blood cells that express the CD34 marker) taken from patients. After growth in culture, the cells were re-infused into the patients. About 20–40% of the re-infused cells had taken up the normal gene. In the patients, this level of gene transfer was adequate, because the treated cells had a robust survival advantage over the mutant cells *in vivo*. So if the rates of cell correction achieved by Urnov *et al.* in human T cells can be achieved in CD34-expressing cells, they will probably be adequate to restore immune function.

Moreover, this approach has the advantage of avoiding unwanted 'integration events', as the strategy is to correct the gene rather than to insert a normal copy. In the earlier study⁷, problems occurred as a consequence of gene integration at an untoward site (next to a growth-promoting gene)⁸. Urnov and

colleagues' approach gets around this problem, offering the possibility of correcting the defect without serious side effects.

The same technique could be applied to the treatment of other diseases in which blood cells are affected, promising safer and more realistic therapies for other forms of SCID and for haemoglobin disorders such as sickle-cell disease. An attractive feature of this particular gene-correction strategy is that it does not require the long-term expression of either the ZFNs or the normal template. So it can be used as a 'hit-and-run' approach. To quote Edward Fitzgerald: "The moving finger writes; and having writ, moves on."

Nevertheless, several hurdles remain before this technique can be successfully translated to humans. For example, it is unclear whether the engineered ZFNs will provoke an immune reaction that rejects these enzymes. And for now, the technique will almost certainly be limited to cell types that can be extracted from the patient, manipulated *ex vivo* and then returned, given that most of the strategies used for *in vivo* gene delivery are too inefficient. Finally, it will be important to determine just how specific the ZFNs are, and whether DNA deletions or rearrangements occur at the target sites. Using Southern blotting, Urnov *et al.* detected neither mistargeting nor gross rearrangements or deletions, but more detailed studies will be required. ■

Katherine A. High is at The Children's Hospital of Philadelphia, 3615 Civic Center Boulevard, Philadelphia, Pennsylvania 19104, USA.
e-mail: high@email.chop.edu

1. Urnov, F. D. *et al.* *Nature* **435**, 646–651 (2005).
2. Porteus, M. H. & Baltimore, D. *Science* **300**, 763 (2003).
3. Bibikova, M. *et al.* *Science* **300**, 764 (2003).
4. Kim, Y. G., Cha, J. & Chandrasegaran, S. *Proc. Natl Acad. Sci. USA* **93**, 1156–1160 (1996).
5. Tüpler, R. *et al.* *Nature* **409**, 832–833 (2001).
6. Klug, A. *FEBS Lett.* **579**, 892–894 (2005).
7. Cavazzana-Calvo, M. *et al.* *Science* **288**, 669–672 (2000).
8. Hacein-Bey-Abina, S. *et al.* *N. Engl. J. Med.* **348**, 255–256 (2003).

GENE REGULATION

Kissing chromosomes

Dimitris Kioussis

A three-dimensional examination of gene regulation suggests that portions from different chromosomes 'communicate' with each other, and bring related genes together in the nucleus to coordinate their expression.

The cell has evolved many strategies to orchestrate gene activation or repression. Spilianakis *et al.*¹ (page 637 of this issue) reveal a novel mechanism of gene regulation, throwing light on how cells organize their genome to respond efficiently to stimuli. They show that genes on different chromosomes that are destined to be expressed within a common cell lineage

are brought together in the nucleus. Such inter-chromosomal communication has been suspected for some time, but this is the first evidence that it actually takes place.

Our understanding of gene regulation has moved from an initial notion of a one-dimensional array of regulatory elements next to each other on the same thread of DNA as the

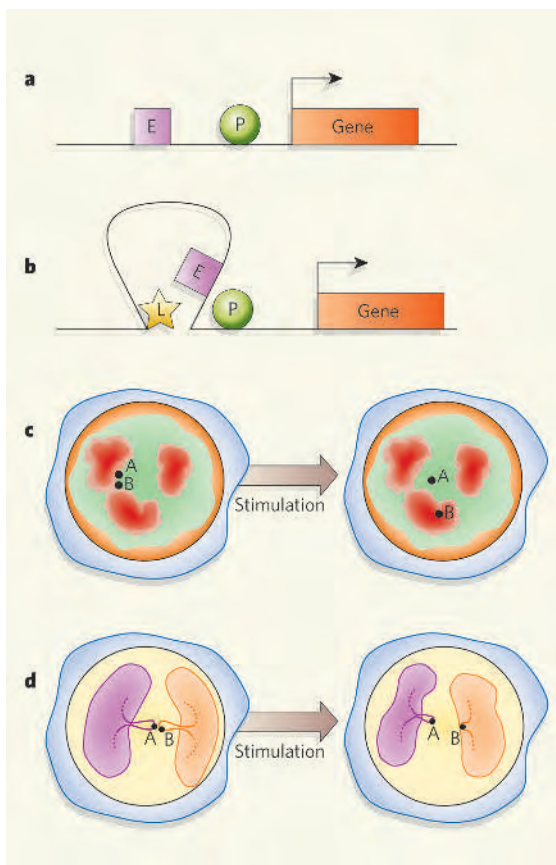


Figure 1 | The three dimensions of gene regulation. **a**, Linear view of gene regulation. The promoter (P) near the start of a gene provides the minimal information needed for gene expression. The function of the promoter is supplemented by enhancers or silencers (E), farther away, where regulatory proteins bind to activate or repress transcription of the gene (arrow). **b**, The 'looping-scanning' model of gene regulation. The locus control region (L) regulates several genes. Proteins binding here scan through large portions of DNA, looping the intervening region out, until they find the relevant gene. **c**, Gene regulation in 3D. Spilianakis *et al.*¹ find that genes from different chromosomes (A and B) are in close proximity until a developmental signal stimulates the cells, when the genes split apart. One moves to a region that represses gene expression (heterochromatin, red) and the other relocates to an area with many active genes (euchromatin, green). **d**, Genes from different chromosomes might come into contact when the chromatin containing them loops out from their chromosome 'territory'.

gene (Fig. 1a) to an appreciation that genes are associated with groups of proteins, forming multimolecular complexes that are arranged in structures generically known as chromatin². The subsequent discovery that distant, contiguous sequences can have a profound effect on gene expression introduced a second dimension onto the scene, with 'looping' and 'scanning' (probably mediated by the attached proteins) invoked to explain these long-range interactions (Fig. 1b)³.

But to explain how genes that are far removed from each other in the genome, and even on different chromosomes, can be coordinated to be expressed together, or to preclude the expression of one another, required a leap into a third dimension. The spatial location of a gene within a cell nucleus can determine whether it is expressed or not: genes residing in areas of chromatin that contain repressive factors (heterochromatin) are silent; conversely, genes in nuclear regions full of activating proteins (euchromatin) are usually switched on (Fig. 1c)^{4,5}. How do genes find their appropriate location in the nucleus of a cell, and how are genes that must be expressed herded into active neighbourhoods?

To address these questions, Spilianakis and colleagues¹ used immune cells called T cells. As they mature, T cells organize themselves into subsets that are assigned specific duties. Thus, T helper (T_H) cells produce factors that help other cells of the immune system to function optimally. After antigen stimulation, naive (undecided) T_H cells develop into either

T_H1 cells, which produce one set of effector molecules (for example interferon (IFN)- γ), or T_H2 cells, which produce a different set (for example interleukin (IL)-4 and IL-5)⁶. The authors explored the organization of two genomic regions within the T_H subsets: the gene encoding IFN- γ (called *Ifng*), which is mainly active in T_H1 cells, and a multi-gene complex including the genes encoding IL-4 and IL-5 (*Il4* and *Il5*), which is mainly active in T_H2 cells.

Viewed through two-dimensional analyses, *Ifng* seems to be regulated by elements found near it on chromosome 10, whereas expression of *Il4* and *Il5* on chromosome 11 seems to be regulated by a 'locus control region' (LCR) on the same chromosome, which directs the entire T_H2 gene complex. Spilianakis *et al.* asked whether these two genetic regions are in close proximity or interact in the nuclei of naive T_H1 or T_H2 cells.

Biochemical and imaging experiments showed that the two regions (on two different chromosomes) are in close proximity in the nucleus of naive T_H cells, but after stimulations that induce a T_H1 or a T_H2 state they seem to move away from each other (Fig. 1c). The authors' interpretation of this is that in the naive T_H cell, the two gene complexes are close together in a region of the nucleus that is poised for gene expression. Upon receiving a specific stimulus, the gene to be activated (for example *Ifng* after a T_H1 stimulus) is allowed to begin expression, whereas its counterpart that is to remain silent (in this case the T_H2 genes)

is moved, presumably to a more repressed region of the nucleus (Fig. 1c).

Spilianakis *et al.* provide some evidence that following stimulation, the inter-chromosomal interactions in the naive cell are replaced by intra-chromosomal ones between regulatory elements within the same gene region. Moreover, the inter-chromosomal association does seem to regulate to some extent the expression of the genes involved, as deletion of a regulatory element (HSS7) from the T_H2 LCR on chromosome 11 disturbs the inter-chromosomal associations, and this is coupled with delayed expression kinetics of *Ifng* from chromosome 10.

These remarkable findings will puzzle us for some time to come. Are inter- and intra-chromosomal associations a general phenomenon occurring in all types of cell? or not⁸? What are the mechanisms that bring two functionally related genes on two different chromosomes together? Between cell divisions, chromosomes expand to occupy 'chromosomal territories' in the nucleus^{9,10}. To allow genes from different chromosomes to come into close proximity, therefore, perhaps chromatin strands might extend from the main body of a chromosome's territory to an area of the nucleus where transcription is possible (Fig. 1d)^{7,11}.

Do both copies of the *Ifng* gene find their functional counterparts of the T_H2 gene complex at some point in the cell's development programme? The results presented by Spilianakis *et al.* are snapshots and indicate that at any particular time the majority of the cells show only one copy of *Ifng* in close proximity to one T_H2 complex. The remaining *Ifng* and T_H2 complex are found separately. Does this mean that there is fluidity in this interaction, allowing genes to come together and then drift apart again after some time? Or does it imply that these genes are expressed from only one of their copies? To start addressing these questions, and others that similar systems may bring up, we would have to develop technologies that allow the detection of inter-chromosomal interactions not as a snapshot, but as a movie. Is it time to go 4D?

Dimitris Kioussis is in the Division of Molecular Immunology, National Institute for Medical Research, The Ridgeway, London NW7 1AA, UK. e-mail: dkioussis@nimr.mrc.ac.uk

- Spilianakis, C. G., Lalioti, M. D., Town, T., Lee, G. R. & Flavell, R. A. *Nature* **435**, 637–645 (2005).
- Felsenfeld, G. & Groudine, M. *Nature* **421**, 448–453 (2003).
- De Laat, W. & Grosfeld, F. *Chromosome Res.* **11**, 447–459 (2003).
- Brown, K. E. *et al. Cell* **3**, 207–217 (1999).
- Kioussis, D. & Festenstein, R. *Curr. Opin. Genet. Dev.* **7**, 614–617 (1997).
- O'Garra, A. *Immunity* **8**, 275–283 (1998).
- Osborne, C. S. *et al. Nature Genet.* **36**, 1065–1071 (2004).
- Kim, S. H. *et al. Cytogenet. Genome Res.* **105**, 292–301 (2004).
- Parada, L. A., Sotiropoulos, S. & Misteli, T. *Exp. Cell Res.* **296**, 64–70 (2004).
- Cremer, T. & Cremer, C. *Nature Rev. Genet.* **2**, 292–301 (2001).
- Kosak, S. T. & Groudine, M. *Genes Dev.* **18**, 1372–1384 (2004).

BRIEF COMMUNICATIONS

A martian meteor and its parent comet

An image of an extraterrestrial meteor was captured as a strange streak in the sky over Mars last year.

Regular meteor showers occur when a planet approaches the orbit of a periodic comet — for example, the Leonid shower is evident around 17 November every year as Earth skims past the dusty trail of comet Tempel–Tuttle. Such showers are expected to occur on Mars as well, and on 7 March last year, the panoramic camera of Spirit, the Mars Exploration Rover, revealed a curious streak across the martian sky. Here we show that the timing and orientation of this streak, and the shape of its light curve, are consistent with the existence of a regular meteor shower associated with the comet Wiseman–Skiff, which could be characterized as martian Cepheids.

On the basis of its orbital elements, comet Wiseman–Skiff is among the top five candidates for producing a regular martian meteor shower^{1,2}. A shower associated with this comet was predicted¹ for 11 March 2004, less than four days after the Spirit picture of the streak (Fig. 1) was taken. As mean regular showers typically last for several days or more, we investigated the possible link between the observed streak and this particular comet.

Meteors from one parent body all seem to emerge from the same point in the sky, called the radiant. Owing to varying velocities in the particle stream, the radiant is not a single point but is typically a few degrees in diameter. The streak seen by Spirit defines a great circle on the sky that passes only 4° from the radiant associated with the orbit of comet Wiseman–Skiff (right ascension, 329.14°; declination, 59.61°; in the constellation Cepheus). More than 96% of all possible radiants produce worse alignments.

This radiant was 10.6° below the horizon, but that still allows the observation of grazing meteors (that is, meteors with a high zenith angle). The radiant could even have been positioned up to 15–20° below the horizon, given the expected range of meteor altitudes (50–100 km; ref. 3) and taking into account the gravitational deviation of the meteoroids from comet Wiseman–Skiff, which travel at relative speeds of 11 km s⁻¹.

The low elevation (14.2°) of the streak, the fact that its trajectory is roughly parallel to the horizon, and its large angular separation from

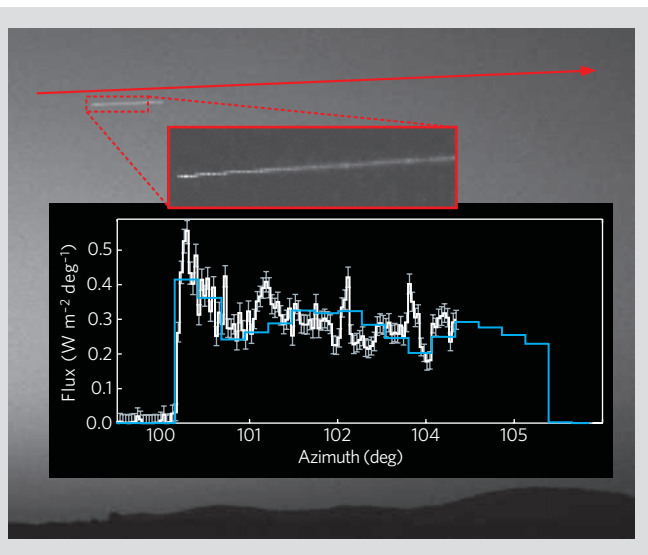


Figure 1 | Image and light curve from the panoramic camera of Spirit, the Mars Exploration Rover.

The streak in the inset image 2P131930937RAD1300P2733L5C2. IMG (ref. 8) is shown in the context of a larger, associated Navcam image. The arrow indicates the predicted direction of a meteor emerging from the predicted radiant in the Cepheus constellation. The plot shows the light curves measured along the streak, taken from the Navcam image (blue curve; each point represents 16 pixels) and from the inset image (white curve with error bars; each point represents 2 pixels). The flux is integrated over a 19-nm spectral band centred at 535 nm.

the radiant (111–115°) are typical features of grazers. We therefore suggest that the streak is a grazing meteor that passed 200–300 km from Spirit, with an observed travel of 13–24 km (4.0° of arc).

The recorded light curve (Fig. 1) is comparable to that of some observed terrestrial meteors^{4,5}. It is characterized by an early peak (assuming that the meteor emerges from Cepheus) and a very sharp initial edge: a 2.5-magnitude change in less than 0.25° (the other edge is truncated by the 15-second exposure). Both effects are expected for high zenith angles^{5,6}, but the abruptness of the edge also requires a slow meteoroid that has a relative speed below 25 km s⁻¹. This is consistent with the 11 km s⁻¹ mean relative speed of particles from comet Wiseman–Skiff.

During the terrestrial 1998 and 2002 Leonids there were sharp peaks of activity, due to Earth's interception of dense swarms of particles ejected by Tempel–Tuttle. Using a model for particle ejection and dynamic evolution⁷, we searched for similar events due to Mars's interception of particle swarms from specific perihelion passages of Wiseman–Skiff. We traced the ejected swarms back to 1900 and found no particular event that could have contributed to a 2004 shower, although there is a promising interception for 20 December 2007. This suggests that the observed meteor belongs to the annual stream responsible for a regular shower, but not to a specific swarm ejected after 1900.

It is therefore likely that we have identified

the first martian meteor and its parent comet. Our findings indicate that martian meteor showers may now be predictable events. Further observations could reveal the chemical effects of meteors on an atmosphere rich in carbon dioxide, a topic pertinent to early Earth's atmospheric chemistry.

Franck Selsis*, **Mark T. Lemmon†**, **Jérémie Vaubaillon‡**, **James F. Bell III§**

*Centre de Recherche Astronomique de Lyon, École Normale Supérieure, 69364 Lyon, France

e-mail: franck.selsis@ens-lyon.fr

†Texas A&M University, Department of Atmospheric Sciences, College Station, Texas 77843-3150, USA

‡IMCCE, Observatoire de Paris, 75014 Paris, France

§Cornell University, Department of Astronomy, Ithaca, New York 14853, USA

1. Selsis, F., Brillet, J. & Rapaport, M. *Astron. Astrophys.* **416**, 783–789 (2004).
2. Christou, A. A. & Beurle, K. *Planet. Space Sci.* **47**, 1475–1485 (1999).
3. Adolfsen, L. G., Gustafson, B. A. S. & Murray, C. D. *Icarus* **119**, 144–152 (1996).
4. Zender, J. J. et al. *Proc. Asteroids, Comets & Meteors ESA-SP 500*, 121–125 (2002).
5. Beech, M. & Murray, I. S. *Mon. Not. R. Astron. Soc.* **345**, 696–704 (2003).
6. Campbell-Brown, M. D. & Koschny, D. *Astron. Astrophys.* **418**, 751–758 (2004).
7. Vaubaillon, J., Colas, F. & Jorda, L. *Astron. Astrophys.* (in the press).
8. Bell, J. et al. *Science* **305**, 800–806 (2004).

doi:435581a

Competing financial interests: declared none.



Cover illustration
Coloured X-ray of a hand with rheumatoid arthritis. The finger joints have become inflamed and swollen. (Du Cane Medical Imaging Ltd/SPL)

Editor, *Nature*
Philip Campbell

Insights Publisher
Sarah Greaves

Insights Editor
Lesley Anson

Production Editors
Maria Hodges
Lucy Odling-Smee

Senior Art Editor
Martin Harrison

Layouts
Nadia Rooney

Diagrams
Nik Spencer

Sponsorship
Claire Hines

Production
Sue Gray

Marketing
Claire Aspinall

Editorial Assistants
Tim Gibbs
Madeline Hutchinson
Laura Shaw

AUTOIMMUNITY

The concept of autoimmunity was first predicted by Nobel Laureate Paul Ehrlich at the start of the twentieth century, and he described it as 'horror autotoxicus'. His experiments led him to conclude that the immune system is normally focused on responding to foreign materials and has an inbuilt tendency to avoid attacking self tissues. But when this process goes wrong, the immune system can attack self tissues resulting in autoimmune disease. The perplexing issue of what allows the immune system to attack self tissues is a continuing focus of research, as the following collection of reviews demonstrates.

In the past, autoimmune diseases have been studied on the basis of the organ affected, but in recent years the focus has switched to a more cross-disciplinary approach with a view to providing a better understanding of the common mechanisms underlying the pathogenesis of these diseases.

This research is now paying off. Previous therapies have essentially been blanket immunosuppressive ones, but recently selective therapies that target pathways common to several autoimmune diseases have been successful in the clinic. These include treatments that target the cytokines tumour necrosis factor and interferon- β . Another approach is the controversial idea of using haematopoietic stem cells for treating severe refractory autoimmune disease.

Such approaches should provide a better understanding of the pathogenic mechanisms of disease and should lead to the development of new therapeutic approaches.

We are pleased to acknowledge the support of Wyeth in producing this Insight. As always, *Nature* carries sole responsibility for all editorial content and peer review.

Elaine Bell, Chief Editor, *Nature Reviews Immunology*
Lucy Bird, Associate Editor, *Nature Reviews Immunology*

REVIEW ARTICLES

584 Paths to understanding the genetic basis of autoimmune disease

J. D. Rioux & A. K. Abbas

590 Cellular and genetic mechanisms of self tolerance and autoimmunity

C. C. Goodnow, J. Sprent,
B. Fazekas de St Groth &
C. G. Vinuesa

598 Regulation of immunity by self-reactive T cells

M. Kronenberg & A. Rudensky

605 An array of possibilities for the study of autoimmunity

C. G. Fathman, L. Soares,
S. M. Chan & P. J. Utz

612 Design of effective immunotherapy for human autoimmunity

M. Feldmann & L. Steinman

620 Treatment of severe autoimmune disease by stem-cell transplantation

M. Sykes & B. Nikolic

nature
insight

Paths to understanding the genetic basis of autoimmune disease

John D. Rioux^{1,2} & Abul K. Abbas³

Some people inherit an unfortunate combination of genetic sequences, such that exposure to an external trigger causes their immune response to turn on their own tissues. Although mutations in a single gene can cause autoimmunity, most autoimmune diseases are associated with several sequence variants. Marked advances in genetic resources and tools are now making it possible to identify the sequence variants that contribute to autoimmune diseases — promising a better understanding of how we normally remain tolerant of our own tissue components, and how this goes wrong in autoimmune disease.

Autoimmune diseases are major causes of morbidity and mortality throughout the world. Many of these diseases tend to be difficult or impossible to cure, for the obvious reason that the focus of the immune response — self antigens — cannot be eliminated. The physical, psychological and economic burden of these diseases is especially devastating because they often attack young adults. The problem is also compounded by the failure of conventional cellular immunological analyses to shed much light on the pathogenic mechanisms. Recently developed therapies, such as tumour necrosis factor (TNF) antagonists, have had some remarkable successes, but these treatments target resulting organ damage and not the (usually unknown) underlying causes. The realization that the development of autoimmunity is strongly influenced by inherited polymorphisms (or DNA sequence variations) brings hope that understanding the genetics of autoimmune diseases will teach us about the causal derangements, and perhaps lead to new therapeutic strategies. Here, we summarize our current understanding of the genetic basis of autoimmunity. We emphasize principles, rather than attempt to list all the reported disease-associated polymorphisms. A brief introduction to the mechanisms of self tolerance and its breakdown provide the foundation for the subsequent discussion of the genetics of autoimmune diseases.

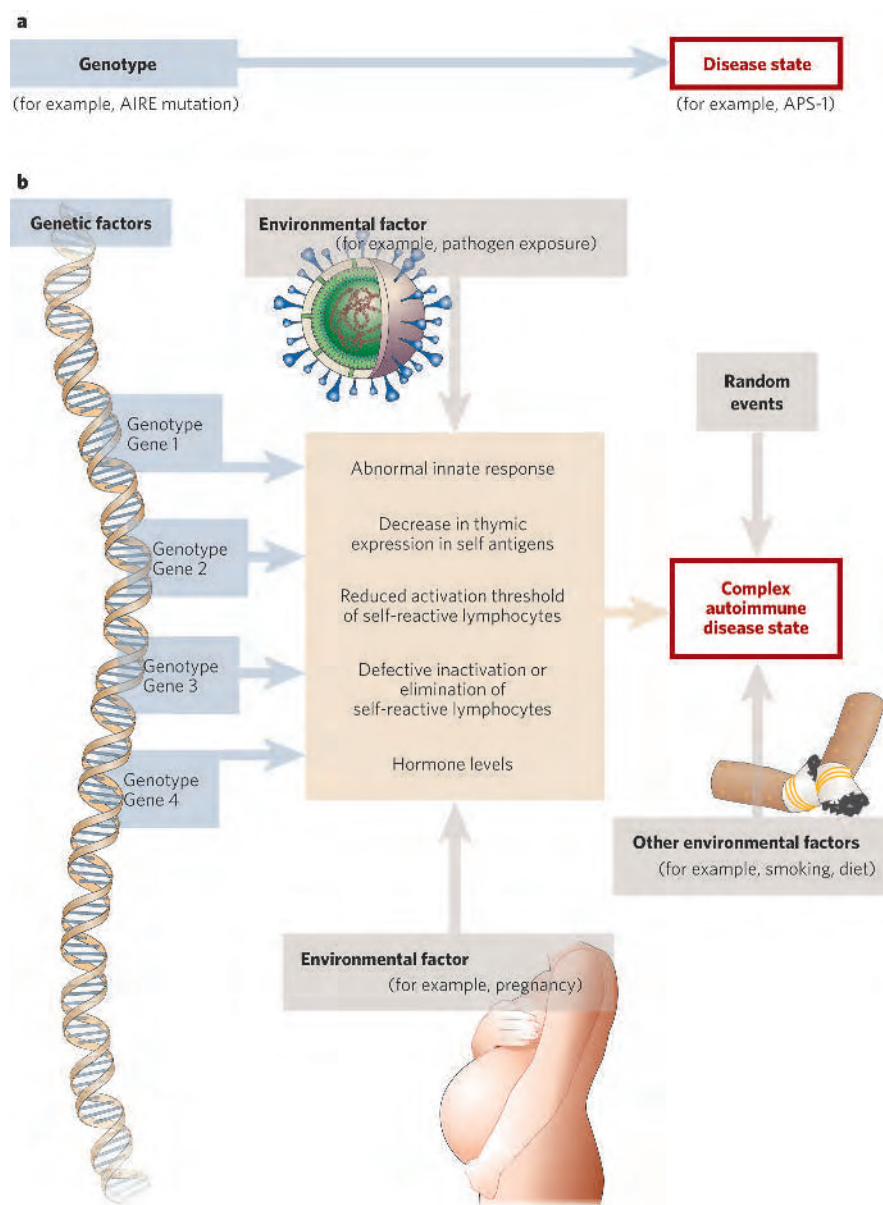
Self tolerance and autoimmunity

All individuals are tolerant of their own potentially antigenic substances, and failure of self tolerance is the fundamental cause of autoimmunity. The mechanisms of self tolerance have been worked out in considerable detail in animal models, and are best understood for CD4⁺ T cells. Self tolerance can be divided into central tolerance and peripheral tolerance. In central tolerance, immature lymphocytes that happen to recognize self antigens in generative lymphoid organs (the bone marrow for B cells and the thymus for T cells) die by apoptosis; in peripheral tolerance, mature self-reactive lymphocytes encounter self antigens in peripheral tissues and are killed or shut off. The principal mechanisms of peripheral tolerance are anergy (functional unresponsiveness), deletion (apoptotic cell death), and suppression by regulatory T cells¹. These mechanisms are described in more detail in the reviews in this issue by Goodnow *et al.* (page 590), and by Kronenberg and Rudensky (page 598). Autoimmune diseases develop when self-reactive lymphocytes escape from tolerance and are activated. Although the mechanisms by which this occurs are not entirely known, autoimmunity is thought to result from a combination of genetic variants, acquired environmental triggers such as infections, and stochastic events.

Table 1 Simple genetic traits associated with autoimmunity				
Gene	Human disease	Mouse mutant or knockout	Mechanism of autoimmunity	References
AIRE	APS-1	Knockout	Decreased expression of self antigens in the thymus, resulting in defective negative selection of self-reactive T cells	2, 3
CTLA4	Association with Graves' disease, type 1 diabetes and others	Knockout	Failure of T cell anergy and reduced activation threshold of self-reactive T cells	9, 65, 66
FOXP3	IPEX (scurfy)	Knockout and mutation	Decreased generation of CD4 ⁺ CD25 ⁺ regulatory T cells	11-13, 67
FAS, FASL	ALPS	lpr/lpr; gld/gld mutants	Failure of apoptotic death of self-reactive B and T cells	16, 68
C4 complement protein	Associated with SLE	Knockout	Defective clearance of immune complexes and possible failure of B cell tolerance	Reviewed in ref. 69

¹Inflammatory Disease Research, The Broad Institute of MIT and Harvard, Harvard Medical School, Brigham and Women's Hospital, 1 Kendall Square, Building 300, Cambridge, Massachusetts 02139-1561, USA (e-mail: rioux@broad.mit.edu). ²Université de Montréal, Montreal Heart Institute, Montreal, Quebec H3C3J7, Canada. ³Department of Pathology, University of California San Francisco, Room M 590, 513 Parnassus Avenue, San Francisco, California 94143, USA.

Figure 1 | Architecture of single-gene disorders versus a model of autoimmune diseases caused by complex traits. **a**, In simple mendelian traits, the relationship between the causal genetic variant (genotype) and the disease state is deterministic. **b**, In complex traits, the clinically recognized disease state results from interactions between multiple genotypes and the environment. Individual genotypes can affect one or more components of the adaptive or innate immune systems; together these lead to an altered immune response to self antigens. On the basis of current findings, the influence of any individual causal allele is modest, and therefore the relationship between the causal variant and the disease state is probabilistic. Although still providing an incomplete picture, the genetic discoveries in mendelian and common diseases are beginning to help build a model of autoimmune disease. The ultimate goal is to build a specific model for each individual disease whereby the effect of individual risk factors (genetic and non-genetic), their interactions, and their impact on disease susceptibility, disease progression and clinical management, are understood.



Genetics of single-gene disorders

To assess the contribution of genetic factors to disease susceptibility, genetic epidemiologists examine the extent of familial clustering; the degree to which monozygotic twins are more concordant for the presence of a disease compared with dizygotic twins; and the increased risk that family members of persons with disease will develop that disease compared with an individual from the general population. Using such estimates of genetic risk, it becomes obvious that in single-gene disorders, the risk conferred on an individual by a given genetic variant is very high, but the overall impact on the population is minimal because these variants are rare.

In these 'simple' diseases (or traits), although there is often variation in phenotypic expression, the relationship between the causal genetic variant and the disease state is deterministic (Fig. 1a). Genome-wide linkage studies attempt to identify genetic markers (and thus a genomic region) where there is more sharing of alleles between individuals (with a given trait) within families than is statistically expected. Such studies are most commonly used in identifying causal genetic variants for single-gene disorders. Genetic studies of simple traits undoubtedly met with earlier success than those of complex traits. However, recent advances in the study of complex traits are now contributing to this knowledge.

Since inborn errors of metabolism were identified as diseases caused

by mutations or deletions affecting single genes, clinicians and scientists have realized that simple genetic traits can teach us a great deal about the pathways of disease, and about normal physiology. In some cases, the genes identified in simple disorders will have more subtle alterations that confer susceptibility to a 'common' disease (that is, diseases caused by a combination of alleles); in other cases, the mutations will simply suggest which biological pathways may be implicated in common diseases. Although many examples exist of single-gene knockout experimental models that have led to autoimmunity, here we focus on genes that are known to be involved in human disorders (Table 1).

AIRE and central tolerance

AIRE (autoimmune regulator) was identified as the gene that is mutated in autoimmune polyendocrine syndrome (APS-1) — a disorder that manifests as autoimmune attack against multiple endocrine organs, the skin and other tissues². In a *tour de force* of functional genomics, the mouse homologue of the gene has been knocked out, and the *AIRE* protein shown to be responsible for the thymic expression of some antigens that are expressed at high levels in different peripheral tissues. In the absence of thymic expression, T cells specific for these antigens escape negative selection (central tolerance), enter the periphery and attack the target tissues^{3,4}. The possibility that thymic defects underlie autoimmunity was raised

long before this elegant demonstration of the importance of AIRE-dependent thymic selection in self tolerance. For example, one of the susceptibility genes associated with type 1 diabetes in humans is insulin: it has been suggested that the disease-associated polymorphism of the insulin gene reduces its expression in the thymus, thus allowing insulin-reactive T cells to escape deletion⁵. It remains to be determined whether defects in central tolerance contribute significantly to most multigenic autoimmune diseases⁶.

CTLA4 and T-cell anergy

Cytotoxic T lymphocyte antigen 4 (CTLA4; CD152) is an inhibitory receptor expressed by T cells that recognizes the costimulatory molecules B7-1 (CD80) and B7-2 (CD86), the ligation of which shuts off T-cell responses and promotes long-lived anergy⁷. CTLA4 works by competitively blocking the engagement of the activating receptor CD28 (by CD80 or CD86), and by transducing inhibitory signals; the latter probably involves tyrosine and serine/threonine phosphatase activation⁸. Germline *CTLA4* knockout mice develop a fatal syndrome of multi-organ lymphocytic infiltrates and severe enlargement of lymphoid organs. Such symptoms are consistent with a systemic autoimmune reaction presumably directed against multiple, as yet unknown, self antigens. This marked demonstration of the obligatory function of CTLA4 has led to a search for polymorphisms in the *CTLA4* gene that are associated with autoimmune diseases. A surprising discovery was that several such diseases, including Graves' disease, type 1 diabetes and other endocrinopathies, show a striking association with a *CTLA4* polymorphism that results in reduced production of a truncated splice variant, which has inhibitory activity⁹. The functional consequences of producing this altered form of CTLA4 have not been defined, but as expected for a complex disease, the biological effect of the causal allele is more subtle than what is found in monogenic disorders or in knockout mice.

FOXP3 and regulatory T cells

FOXP3 (encoding a transcription factor of the forkhead family) is a striking example of a gene whose role in autoimmunity has been revealed by the confluence of animal studies and studies of a quite rare human disease. CD4⁺CD25⁺ regulatory T cells, now established as major controllers of immune responses to self and other antigens¹⁰, were shown to express high levels of *FOXP3*. Three groups demonstrated that induced knockout or spontaneous mutation of the mouse *Foxp3* gene ('scurfy' mice) led to a systemic autoimmune disease associated with the absence of CD4⁺CD25⁺ regulatory T cells^{11–13}. At the same time, a human disease known by the acronym IPEX (immune dysregulation, polyendocrinopathy, enteropathy, X-linked syndrome) was shown to be associated with mutations in *FOXP3*. These results indicate that the development and/or function of regulatory T cells are dependent on the activity of this one transcription factor; its downstream targets and major functions are not yet defined.

FAS and lymphocyte apoptosis

Fas (CD95) is the prototype of a death receptor of the TNF receptor family. Its biological importance was established by seminal studies

demonstrating that in two mouse models of autoimmunity and lymphoproliferation (named the *lpr/lpr* and *gld/gld* strains), genetic lesions affected the genes encoding Fas and Fas ligand, respectively¹⁴. These results were among the first to show that a single genetic abnormality could give rise to a complex autoimmune phenotype, and that the failure of one apoptosis pathway resulted in self-tolerance breakdown. The Fas death receptor contributes to the deletion of mature T and B cells that recognize self antigens¹⁵. A human disease, called autoimmune lymphoproliferative syndrome (ALPS), which resembles the disease of *lpr* and *gld* mice, is caused by *Fas* mutations¹⁶. Thus, this is an excellent example of how mouse models can lead to a better understanding of human diseases. Although these autoimmune diseases have a superficial similarity to human systemic lupus erythematosus (SLE), there is no convincing evidence for *FAS* polymorphisms or mutations being associated with SLE.

The elegant simplicity of these monogenic disorders has been key to using them to elucidate mechanisms of self tolerance and autoimmunity. The experimental models also illustrate how animal studies inform analyses of human disease, and the striking influences of 'background' genes on the severity and manifestations of diseases. However, as stated at the outset, these single-gene models of autoimmunity are rare; most autoimmune diseases are complex, multigenic traits. The remainder of our discussion will focus on these complex disorders.

Genetics of common autoimmune diseases

In contrast to simple diseases, common diseases are believed to result from a combination of susceptibility alleles at multiple loci, environmental factors (such as smoking, pathogen exposure and hormone levels), and stochastic events (Fig. 1b). There is some debate as to whether common diseases are caused by common alleles of low penetrance (that is, alleles that confer modest increased risk to disease) or by multiple rare alleles of high penetrance. Although most disease alleles identified for common autoimmune diseases have the former characteristics (Table 2), these may not represent the full spectrum of disease alleles. Our ability to discover as yet unidentified disease alleles has been improved by the recent explosion of knowledge regarding the common genetic variation in the human genome. Specifically, recent work has demonstrated that the vast majority of genetic variation in the human genome consists of individual bases that exist as either of two alleles in the population — known as single nucleotide polymorphisms (SNPs) — rather than as deletions or rearrangements. Approximately ten million SNPs in the human genome have a minor allele frequency greater than 1% and represent about 90% of the genetic variation in the human genome. Initial efforts to discover and map SNPs to the reference sequence of the human genome have resulted in a public resource containing most of these common SNPs^{17–19}.

Over the past decade, in autoimmunity as in many other human diseases, there has been great interest in testing candidate genomic regions (identified by linkage studies), or candidate genes (selected on the basis of their location under a linkage peak, or their known functional properties, or both) for evidence of their association with disease. Before the creation of public SNP databases, sequencing was

Table 2 Relative risk, frequency and population attributable risk for a set of causal variants						
Variant	CARD15*	IBD5	PTPN22†	PTPN22‡	CTLA4§	CTLA4§
Disease	Crohn's disease	Crohn's disease	Type 1 diabetes	Rheumatoid arthritis	Type 1 diabetes	Graves' disease
Relative risk	3	1.3	1.7	1.65	1.14	1.5
(95% CI)		(1.2–1.4)	(1.5–1.9)	(1.2–2.2)	(1.1–1.2)	(1.3–1.75)
Frequency	~7%	~35%	~10%	~10%	~50%	~50%
PAR	23%	18%	16%	~12%	13%	36%
References	30	56	37	39	9	9
The genetic characteristics of a selected set of susceptibility genes identified for common autoimmune diseases. The finding of all these loci has been replicated in multiple studies. Therefore, accurate estimates are provided of the risk of disease that each locus confers. Most of these susceptibility loci confer modest risk (1.1–1.7-fold increased risk to carriers), but given their relatively common frequency in the population (7–50%), have a significant impact on disease burden — as estimated, for example, by the population-attributable risk (PAR). (In contrast, the PAR for causal genes in simple disorders is expected to approach 100%). The modest risk conferred by a given locus in common autoimmune diseases does not provide any significant statistical power to predict whether a given person will develop a disease but may be useful in helping to classify patients into molecularly relevant subgroups. CI, confidence interval; *Calculated for heterozygotes for any of the three characterized mutations (Arg702Trp, Gly908Arg and Leu1007fsinsC); †Calculated for the 620W allele; ‡Calculated for the CT60/G allele.						

needed to identify the genetic variants to test in an association study. Now it seems that there is almost an excess of SNPs to type: do we need to type all ten million to find the few that are causing a given autoimmune disease? To answer this question, it is important to understand the relationships that exist between these SNPs. First, by examining a high density of specific areas of the genome^{20–22}, and by performing genome-wide surveys²³, it has become clear that the alleles of the SNPs form patterns (also known as haplotypes) in the genome. Specifically, the emerging data demonstrate that alleles at nearby SNPs are highly correlated with one another (this is known as linkage disequilibrium). The tight correlation between SNPs, however, is broken down by recombination events resulting in haplotype blocks. Furthermore, the existing data and models based on these data suggest that, rather than occurring randomly, there are recombination hotspots in the genome^{22,24,25}. Knowledge of the haplotype structure not only allows an optimal subset of SNPs to be selected that efficiently extracts the information about the common patterns of variation, but it also can direct how such data should be analysed. For these reasons, an international effort is underway to map the common patterns of genetic variation across the entire genome, and it is expected that the first phase will be completed by mid-2005 (ref. 26). Already the information from this International HapMap can be incorporated into the design of association studies of candidate genes/genomic regions, and it will soon be possible to apply this new information to the design and execution of powerful genome-wide association studies^{27,28}.

Successes in mapping susceptibility genes

Given the early lack of success in identifying genes for complex traits, it was believed that successes in human disease genetics might be limited to disorders with single-gene mendelian inheritance. However, this changed with the publication of several studies reporting the mapping of susceptibility loci in complex human traits: many of these studies took advantage of the haplotype structure of the genome to narrow down the region of association. Early in 2000, fine-scale mapping of a complex human trait successfully led to the localization of the type 1 diabetes locus in the major histocompatibility complex (MHC) to a discrete 570-kilobase (kb) region²⁹. An association mapping approach also proved successful when Hugot and colleagues identified sequence variants in the *NOD2* (*CARD15*) gene that are associated with Crohn's disease³⁰. Comprehensive association mapping approaches successfully identified haplotypic variation in the cytokine gene cluster on chromosome 5q31, which confers susceptibility to Crohn's disease³¹, for which two potentially causal variants have recently been proposed³². The *ADAM33* and *GPRA* susceptibility genes for asthma³³ have similarly been identified. Large association studies of candidate genes have also resulted in important discoveries: the *IDDM12/CTLA4* locus in Graves' disease and in type 1 diabetes⁹; the *NOD2* gene in Crohn's disease^{34,35}; the *PTPN22* gene in type 1 diabetes, rheumatoid arthritis and SLE^{36–39}; and the *PDCD1* gene in SLE and rheumatoid arthritis^{40,41}. The observation that genes such as *CTLA4* and *PTPN22* are associated with multiple disorders is consistent with the hypothesis that certain immunological pathways are common to multiple autoimmune diseases, whereas other pathophysiological mechanisms are specific to a particular disease.

The identification of these susceptibility genes provides the immunology community with an opportunity to study the key pathways and molecular mechanisms that can lead to increased disease susceptibility. Initial functional studies have begun to provide some relevant clues. For example, work regarding the effect of genetic variants in *CTLA4* on susceptibility to type 1 diabetes in mice and humans has provided substantial evidence that modulating the expression of alternative splice forms can regulate the expression of T-cell-mediated autoimmunity^{9,42}. Specifically, the human variant leads to the decreased production of a ligand-independent form of CTLA4 — a splice form believed to be involved in the downregulation of memory/effector T-cell activation^{9,42}. Genetic variation in the human *PTPN22* gene also seems to modulate T-cell activity. However, this is

due to an amino acid change in the encoded protein, the lymphoid protein tyrosine phosphatase (LYP), a known suppressor of T-cell activation. Bottini and colleagues⁴³ hypothesized that *PTPN22* was a good candidate gene for autoimmunity because it and other protein tyrosine phosphatases are involved in preventing spontaneous T-cell activation. These investigators identified an SNP that was associated with type 1 diabetes⁴³, which changed an amino acid involved in the interaction between LYP and the negative regulatory kinase CSK. *PDCD1*, a positional and functional candidate gene, encodes the programmed cell death 1 gene whose gene product has been shown to regulate peripheral tolerance in T and B cells^{44,45}. An intronic variant in the *PDCD1* gene was proposed to lead to aberrant *PDCD1* regulation and increase an individual's susceptibility to SLE⁴¹. These few examples demonstrate how variants that modify gene expression or protein structure can lead to modulation of the adaptive immune responses.

Alteration of the innate immune system, on the other hand, may be the primary effect of the two missense mutations and the truncation mutant in the *NOD2* gene, all of which have been associated with Crohn's disease. Specifically, the *NOD2* gene product has been shown to recognize muramyl dipeptide (MDP) — a component of the bacterial wall. The Crohn's disease-associated mutations were believed to lead to impaired nuclear factor- κ B (NF- κ B) signalling following recognition of MDP^{34,46,47}. Using spleen macrophages derived from *Nod2*-deficient mice, Watanabe and colleagues proposed an alternative explanation. Specifically, they suggested that the *NOD2* gene product limits the pro-inflammatory effects mediated by Toll-like receptor 2 (TLR2) signalling, such that mutations in *NOD2* would lead to excessive T-helper type 1 (T_H1) responses^{46,47}. Efforts to knock out the *Nod2* gene have not, however, led to gastrointestinal inflammation. This perhaps is not surprising given the complexity of the human disease; no single genetic or environmental factor is expected to be necessary or sufficient to cause the disease. These *Nod2*^{-/-} mice also did not show differential susceptibility to chemically induced gastrointestinal inflammation, but they did seem to be more susceptible to oral challenge with Gram-positive bacteria^{48,49}. The results from knocking-in the human frameshift mutation into mice (*Nod2*^{2939C}), however, seem to suggest a very different mechanism: as opposed to the lack of response to MDP in the *Nod2*^{-/-} mice, the *Nod2*^{2939C} mice have an elevated response to MDP and a heightened intestinal inflammatory response to chemical injury⁵⁰.

These exciting experiments demonstrate the importance of a genetic and environmental context for such functional studies and the challenges that lie ahead in deciphering the exact mechanisms by which genetic variation can lead to autoimmunity. Although it is still too early to know the precise mechanisms by which individual allelic variants confer susceptibility, the elucidation of the pathogenic mechanisms in the coming years will undoubtedly shed much light on the common and disease-specific processes.

Challenges in mapping susceptibility genes

Genetic association studies of autoimmune diseases, as for all other complex human traits, face two main challenges: the ability to distinguish between true and false associations, and the ability to demonstrate causality. The first challenge relates to the nature of many causal alleles in common disease; many of the implicated alleles only confer a modest increased risk of developing disease (see Table 2). The power to detect disease susceptibility genes is influenced by the magnitude of the risk conferred by the susceptibility allele, and by its frequency in the population. The weaker the effect of an allele, the greater the sample size required to detect an association signal that can be distinguished from the background noise of an association study. Unfortunately, it is not possible to know *a priori* the frequency and the strength of disease alleles. However, the detection of modest alleles may require many thousands of samples²⁸. This is further complicated by the fact that the first report of an association often overestimates the effect of the putative associated allele⁵¹. This effect is known as the 'winner's curse' — a mathematical concept first described in the context of auctions and bilateral negotiations⁵². Replication studies, there-

fore, often produce seemingly inconsistent results when compared with the original study^{53,54}. Luckily, this challenge can be addressed by performing association studies (novel and replication) with large sample sizes, and by combining the results from multiple studies using a meta-analysis approach^{53,55}. Indeed, these two approaches have successfully demonstrated that a number of susceptibility genes for autoimmune disease (for example, *PTPN22* in type 1 diabetes, *IBD5* and *NOD2* in Crohn's disease) can be convincingly identified and replicated^{37,56,57}.

The second challenge relates to the patterns of linkage disequilibrium described above. Although intervals of historical recombination occur between haplotype blocks, this is a modest enough effect for correlations between blocks to persist²¹. Given the extreme local variation in the magnitude of this recombination²⁵ and in gene density¹⁹, genomic regions implicated by association studies may contain a single gene or multiple genes. For the latter, current sample sizes may be insufficient to narrow down the implicated region to less than a handful of genes. One example is a 250-kb haplotype on chromosome 5q31 that was found to be associated with Crohn's disease³¹. The haplotype associated with increased risk of developing Crohn's disease extends uninterrupted across multiple blocks, as the recombination events that have occurred in the intervals in this region have primarily involved the haplotypes that do not confer risk. Therefore, this association implicates a region that contains a minimum of four genes (*IRF1*, *SLC22A5* (*OCTN2*), *SLC22A4* (*OCTN1*) and *PDLIM3*) and multiple SNPs that have an allele that is unique to the risk haplotype. Therefore, the identification of the causal gene(s) and related variant(s) will rely on functional studies that link gene/variants and phenotype. Another important example is that of the MHC, where extended haplotypes are believed to represent a common feature of this region⁵⁸. Importantly, the MHC region has been associated with almost all autoimmune diseases and contains the HLA genes: HLA gene products are crucial for antigen presentation to cells of the adaptive immune system and for controlling some reactions of the innate immune system, such as activation of natural killer (NK) cells. This genomic region also contains hundreds of other genes, many with putative or proven function in the immune system⁵⁹. Most association studies of this genomic region, however, have been limited to one or a few of the HLA genes. The preliminary efforts to map the patterns of genetic variation in this region suggest that dense sets of SNPs applied to large study cohorts will enable the identification of the HLA and non-HLA components of the genetic susceptibility conferred by the MHC region^{60–63}.

Future prospects

One of the great promises of genetics is to help build a molecular model of disease that combines information regarding genotype and environment, as well as their various interactions (Fig. 1b). One challenge that is specific to immune-related disorders is the fact that an individual's functional repertoire of specific antigen receptors (B-cell receptors and T-cell receptors) is a product of the interaction between their genetic repertoire and the environment. It will therefore be important to define an individual's functional state in the context of their genetic background. Despite this challenge, some of the discoveries so far in single-gene and complex forms of autoimmunity are helping to build such a model. These models could potentially help to assess an individual's risk of developing disease. There is also some indication that this information will be relevant for assigning patients to molecular subgroups of these very heterogeneous diseases, which may help to predict a particular disease outcome and/or response to therapy.

A complementary approach to improving the response to current therapies is to examine variation in the genes controlling the absorption, distribution, metabolism and excretion of the drugs used to dampen the chronic inflammatory response in patients⁶⁴. An aspect of these models that may prove to be important is knowing which genes are common to more than one autoimmune disease versus those that are disease-specific, as these might turn out to be qualitatively very different drug targets. Finally, there is no doubt that any novel suscepti-

bility gene, or even newly identified causal variation in a previously well characterized gene, will provide the immunologist with some very interesting avenues to follow in their path to understanding the fundamental mechanisms of autoimmunity. ■

- Walker, L. S. & Abbas, A. K. The enemy within: keeping self-reactive T cells at bay in the periphery. *Nature Rev. Immunol.* **2**, 11–19 (2002).
- Björns, P., Aaltonen, J., Horelli-Kuitunen, N., Yaspo, M. L. & Peltonen, L. Gene defect behind APECED: a new clue to autoimmunity. *Hum. Mol. Genet.* **7**, 1547–1553 (1998).
- Anderson, M. S. *et al.* Projection of an immunological self shadow within the thymus by the aire protein. *Science* **298**, 1395–1401 (2002).
- Liston, A., Lesage, S., Wilson, J., Peltonen, L. & Goodnow, C. C. Aire regulates negative selection of organ-specific T cells. *Nature Immunol.* **4**, 350–354 (2003).
- Eisenbarth, G. S. Insulin autoimmunity: immunogenetics/immunopathogenesis of type 1A diabetes. *Ann. NY Acad. Sci.* **1005**, 109–118 (2003).
- Mathis, D. & Benoist, C. Back to central tolerance. *Immunity* **20**, 509–516 (2004).
- Salomon, B. & Bluestone, J. A. Complexities of CD28/B7: CTLA-4 costimulatory pathways in autoimmunity and transplantation. *Annu. Rev. Immunol.* **19**, 225–252 (2001).
- Baroja, M. L. *et al.* Inhibition of CTLA-4 function by the regulatory subunit of serine/threonine phosphatase 2A. *J. Immunol.* **168**, 5070–5078 (2002).
- Ueda, H. *et al.* Association of the T-cell regulatory gene CTLA4 with susceptibility to autoimmune disease. *Nature* **423**, 506–511 (2003).
- Sakaguchi, S. Naturally arising CD4⁺ regulatory T cells for immunologic self-tolerance and negative control of immune responses. *Annu. Rev. Immunol.* **22**, 531–562 (2004).
- Hori, S., Nomura, T. & Sakaguchi, S. Control of regulatory T cell development by the transcription factor Foxp3. *Science* **299**, 1057–1061 (2003).
- Fontenot, J. D., Gavin, M. A. & Rudensky, A. Y. Foxp3 programs the development and function of CD4⁺CD25⁺ regulatory T cells. *Nature Immunol.* **4**, 330–336 (2003).
- Khattari, R., Cox, T., Yasayko, S. A. & Ramsdell, F. An essential role for Scurfin in CD4⁺CD25⁺ T regulatory cells. *Nature Immunol.* **4**, 337–342 (2003).
- Nagata, S. & Suda, T. Fas and Fas ligand: lpr and gld mutations. *Immunol. Today* **16**, 39–43 (1995).
- Lenardo, M. *et al.* Mature T lymphocyte apoptosis—immune regulation in a dynamic and unpredictable antigenic environment. *Annu. Rev. Immunol.* **17**, 221–253 (1999).
- Fischer, A., Rieux-Laucat, F. & Le Deist, F. Autoimmune lymphoproliferative syndromes (ALPS): models for the study of peripheral tolerance. *Rev. Immunogenet.* **2**, 52–60 (2000).
- Sachidanandam, R. *et al.* A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**, 928–933 (2001).
- Wang, D. G. *et al.* Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* **280**, 1077–1082 (1998).
- International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945 (2004).
- Patil, N. *et al.* Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* **294**, 1719–1723 (2001).
- Daly, M. J., Rioux, J. D., Schaffner, S. F., Hudson, T. J. & Lander, E. S. High-resolution haplotype structure in the human genome. *Nature Genet.* **29**, 229–232 (2001).
- Jeffreys, A. J., Kauppi, L. & Neumann, R. Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nature Genet.* **29**, 217–222 (2001).
- Gabriel, S. B. *et al.* The structure of haplotype blocks in the human genome. *Science* **296**, 2225–2229 (2002).
- Reich, D. E. *et al.* Human genome sequence variation and the influence of gene history, mutation and recombination. *Nature Genet.* **32**, 135–142 (2002).
- McVean, G. A. *et al.* The fine-scale structure of recombination rate variation in the human genome. *Science* **304**, 581–584 (2004).
- The International HapMap Consortium. The International HapMap Project. *Nature* **426**, 789–796 (2003).
- Hirschhorn, J. N. & Daly, M. J. Genome-wide association studies for common diseases and complex traits. *Nature Rev. Genet.* **6**, 95–108 (2005).
- Wang, W. Y., Barratt, B. J., Clayton, D. G. & Todd, J. A. Genome-wide association studies: theoretical and practical concerns. *Nature Rev. Genet.* **6**, 109–118 (2005).
- Herr, M. *et al.* Evaluation of fine mapping strategies for a multifactorial disease locus: systematic linkage and association analysis of IDDM1 in the HLA region on chromosome 6p21. *Hum. Mol. Genet.* **9**, 1291–1301 (2000).
- Hugot, J. P. *et al.* Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature* **411**, 599–603 (2001).
- Rioux, J. D. *et al.* Genetic variation in the 5q31 cytokine gene cluster confers susceptibility to Crohn disease. *Nature Genet.* **29**, 223–228 (2001).
- Pelteková, V. D. *et al.* Functional variants of OCTN cation transporter genes are associated with Crohn disease. *Nature Genet.* **36**, 471–475 (2004).
- Laitinen, T. *et al.* Characterization of a common susceptibility locus for asthma-related traits. *Science* **304**, 300–304 (2004).
- Ogura, Y. *et al.* A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease. *Nature* **411**, 603–606 (2001).
- Hampe, J. *et al.* Association between insertion mutation in NOD2 gene and Crohn's disease in German and British populations. *Lancet* **357**, 1925–1928 (2001).
- Onengut-Gumuscu, S., Ewens, K. G., Spielman, R. S. & Concannon, P. A functional polymorphism (1858C/T) in the *PTPN22* gene is linked and associated with type 1 diabetes in multiplex families. *Genes Immun.* **5**, 678–680 (2004).
- Smyth, D. *et al.* Replication of an association between the lymphoid tyrosine phosphatase locus (LYP/*PTPN22*) with type 1 diabetes, and evidence for its role as a general autoimmunity locus. *Diabetes* **53**, 3020–3023 (2004).
- Kyogoku, C. *et al.* Genetic association of the R620W polymorphism of protein tyrosine phosphatase PTPN22 with human SLE. *Am. J. Hum. Genet.* **75**, 504–507 (2004).
- Begovich, A. B. *et al.* A missense single-nucleotide polymorphism in a gene encoding a protein tyrosine phosphatase (*PTPN22*) is associated with rheumatoid arthritis. *Am. J. Hum. Genet.* **75**, 330–337 (2004).

40. Lin, S. C. *et al.* Association of a programmed death 1 gene polymorphism with the development of rheumatoid arthritis, but not systemic lupus erythematosus. *Arthritis Rheum.* **50**, 770–775 (2004).
41. Prokunina, L. *et al.* A regulatory polymorphism in *PDCD1* is associated with susceptibility to systemic lupus erythematosus in humans. *Nature Genet.* **32**, 666–669 (2002).
42. Vijaykrishnan, L. *et al.* An autoimmune disease-associated CTLA-4 splice variant lacking the B7 binding domain signals negatively in T cells. *Immunity* **20**, 563–575 (2004).
43. Bottini, N. *et al.* A functional variant of lymphoid tyrosine phosphatase is associated with type 1 diabetes. *Nature Genet.* **36**, 337–338 (2004).
44. Nishimura, H. & Honjo, T. PD-1: an inhibitory immunoreceptor involved in peripheral tolerance. *Trends Immunol.* **22**, 265–268 (2001).
45. Blank, C. *et al.* Absence of programmed death receptor 1 alters thymic development and enhances generation of CD4/CD8 double-negative TCR-transgenic T cells. *J. Immunol.* **171**, 4574–4581 (2003).
46. Watanabe, T., Kitani, A., Murray, P. J. & Strober, W. NOD2 is a negative regulator of Toll-like receptor 2-mediated T helper type 1 responses. *Nature Immunol.* **5**, 800–808 (2004).
47. Girardin, S. E. *et al.* Nod2 is a general sensor of peptidoglycan through muramyl dipeptide (MDP) detection. *J. Biol. Chem.* **278**, 8869–8872 (2003).
48. Pauleau, A. L. & Murray, P. J. Role of Nod2 in the response of macrophages to Toll-like receptor agonists. *Mol. Cell. Biol.* **23**, 7531–7539 (2003).
49. Kobayashi, K. S. *et al.* Nod2-dependent regulation of innate and adaptive immunity in the intestinal tract. *Science* **307**, 731–734 (2005).
50. Maeda, S. *et al.* Nod2 mutation in Crohn's disease potentiates NF- κ B activity and IL-1 β processing. *Science* **307**, 734–738 (2005).
51. Ioannidis, J. P., Ntzani, E. E., Trikalinos, T. A. & Contopoulos-Ioannidis, D. G. Replication validity of genetic association studies. *Nature Genet.* **29**, 306–309 (2001).
52. Bazerman, M. & Samuelson, W. The Winners Curse — an Empirical Investigation. *Lecture Notes in Economics and Mathematical Systems* **213**, 186–200 (1983).
53. Lohmueller, K. E., Pearce, C. L., Pike, M., Lander, E. S. & Hirschhorn, J. N. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nature Genet.* **33**, 177–182 (2003).
54. Hirschhorn, J. N., Lohmueller, K., Byrne, E. & Hirschhorn, K. A comprehensive review of genetic association studies. *Genet. Med.* **4**, 45–61 (2002).
55. Ioannidis, J. P., Trikalinos, T. A., Ntzani, E. E. & Contopoulos-Ioannidis, D. G. Genetic associations in large versus small studies: an empirical assessment. *Lancet* **361**, 567–571 (2003).
56. Daly, M. J. & Rioux, J. D. New approaches to gene hunting in IBD. *Inflamm. Bowel Dis.* **10**, 312–317 (2004).
57. Hugot, J. P. Genetic origin of IBD. *Inflamm. Bowel Dis.* **10**, (Suppl. 1) S11–S15 (2004).
58. Alper, C. A., Awdeh, Z. & Yunis, E. J. Conserved, extended MHC haplotypes. *Exp. Clin. Immunogenet.* **9**, 58–71 (1992).
59. Horton, R. *et al.* Gene map of the extended human MHC. *Nature Rev. Genet.* **5**, 889–899 (2004).
60. Stenzel, A. *et al.* Patterns of linkage disequilibrium in the MHC region on human chromosome 6p. *Hum. Genet.* **114**, 377–385 (2004).
61. Walsh, E. C. *et al.* An integrated haplotype map of the human major histocompatibility complex. *Am. J. Hum. Genet.* **73**, 580–590 (2003).
62. Ahmad, T. *et al.* Haplotype-specific linkage disequilibrium patterns define the genetic topography of the human MHC. *Hum. Mol. Genet.* **12**, 647–656 (2003).
63. Miretti, M. M. *et al.* A high-resolution linkage-disequilibrium map of the human major histocompatibility complex and first generation of tag single-nucleotide polymorphisms. *Am. J. Hum. Genet.* **76**, 634–646 (2005).
64. Ahmadi, K. R. *et al.* A single-nucleotide polymorphism tagging set for human drug metabolism and transport. *Nature Genet.* **37**, 84–89 (2005).
65. Tivol, E. A. *et al.* Loss of CTLA-4 leads to massive lymphoproliferation and fatal multiorgan tissue destruction, revealing a critical negative regulatory role of CTLA-4. *Immunity* **3**, 541–547 (1995).
66. Waterhouse, P. *et al.* Lymphoproliferative disorders with early lethality in mice deficient in *Ctla-4*. *Science* **270**, 985–988 (1995).
67. Bennett, C. L. *et al.* The immune dysregulation, polyendocrinopathy, enteropathy, X-linked syndrome (IPEX) is caused by mutations of *FOXP3*. *Nature Genet.* **27**, 20–21 (2001).
68. Fisher, G. H. *et al.* Dominant interfering *Fas* gene mutations impair apoptosis in a human autoimmune lymphoproliferative syndrome. *Cell* **81**, 935–946 (1995).
69. Atkinson, J. P. Complement deficiency: predisposing factor to autoimmune syndromes. *Clin. Exp. Rheumatol.* **7**, (Suppl. 3) S95–S101 (1989).

Acknowledgements We thank J. Bluestone, M. Anderson, M. Daly and F. Sanchez, for their helpful comments. We also thank L. Gaffney for her help in the preparation of the manuscript. J.D.R. is supported by grants from the CCFA and NIDDK.

Competing interests statement The authors declare that they have no competing financial interests.

Cellular and genetic mechanisms of self tolerance and autoimmunity

Christopher C. Goodnow^{1,2}, Jonathon Sprent³, Barbara Fazekas de St Groth⁴ & Carola G. Vinuesa¹

The mammalian immune system has an extraordinary potential for making receptors that sense and neutralize any chemical entity entering the body. Inevitably, some of these receptors recognize components of our own body, and so cellular mechanisms have evolved to control the activity of these 'forbidden' receptors and achieve immunological self tolerance. Many of the genes and proteins involved are conserved between humans and other mammals. This provides the bridge between clinical studies and mechanisms defined in experimental animals to understand how sets of gene products coordinate self-tolerance mechanisms and how defects in these controls lead to autoimmune disease.

Our immune system is the body's sixth sense. It can react to any chemical structure imaginable to fight off every possible microorganism. The receptors coordinating this feat are antibodies expressed on the surface of B cells as B-cell receptors (BCRs) and T-cell receptors (TCRs) displayed on T cells. Huge receptor diversity is encoded in the mammalian genome by two processes of somatic genome modification that occur selectively in lymphocytes. First, V(D)J recombination assembles unique BCR and TCR genes from three separate gene segments, the variable (V), diversity (D) and joining (J) genes, during B-

and T-cell differentiation. This takes place in the 'central lymphoid tissues', which are principally the bone marrow for B cells and the thymus for T cells. Second, somatic hypermutation substitutes single nucleotides of BCR genes during a late phase of the immune response in peripheral lymphoid tissues (such as the spleen, lymph nodes and tonsils). A significant fraction of the receptors generated by both these processes bind to one or more self components in the body — a by-product of a deliberately random receptor-generating process. Between 20 and 50% of TCRs and BCRs generated by V(D)J recombination

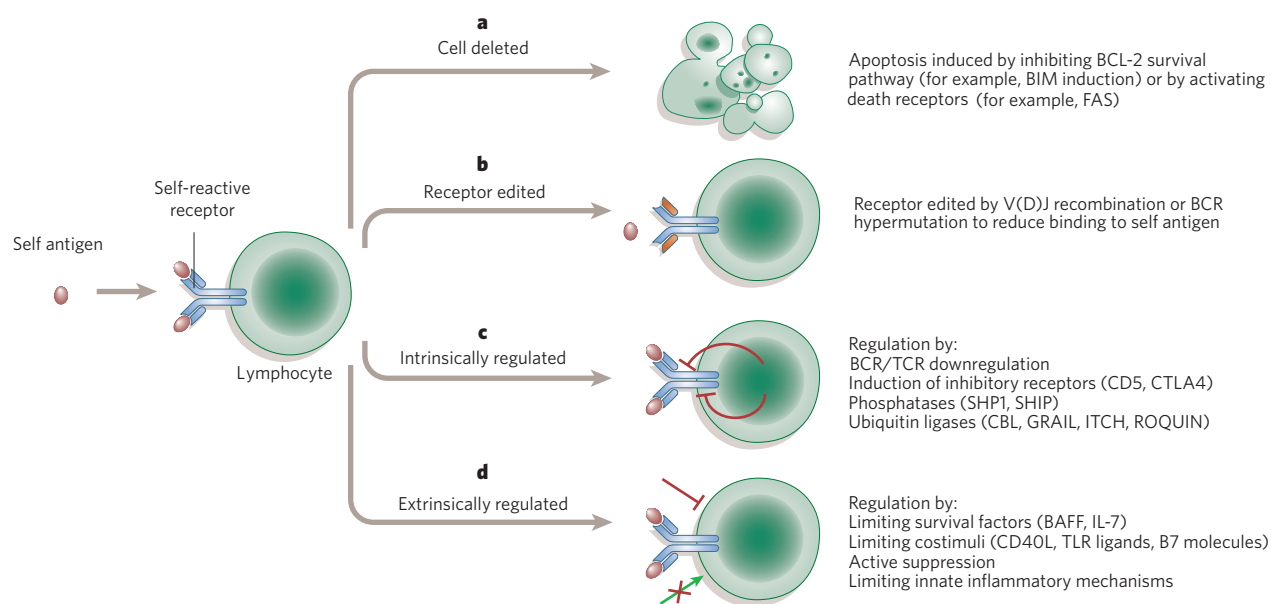


Figure 1 | Four cellular strategies are used to regulate self-reactive receptors at different points during B- and T-cell differentiation. a, The cell is deleted through induction of cell death. **b**, The receptor is edited to one that is less self-reactive. **c**, Biochemical or gene-expression changes intrinsically

dampen the self-reactive receptor's ability to activate the cell. **d**, The ability of self-reactive cells or antibody to cause autoimmunity is limited by using extrinsic suppression and by limiting essential growth factors, costimuli and inflammatory mediators.

¹John Curtin School of Medical Research and ²Australian Phenomics Facility, The Australian National University, Canberra ACT 2601 Australia (e-mail: Chris.Goodnow@anu.edu.au).

³Department of Immunology, IMM4, The Scripps Research Institute, 10550 N. Torrey Pines Road, La Jolla, California 92037, USA. ⁴Centenary Institute of Cancer Medicine and Cell Biology, Newtown, NSW 2042, Australia

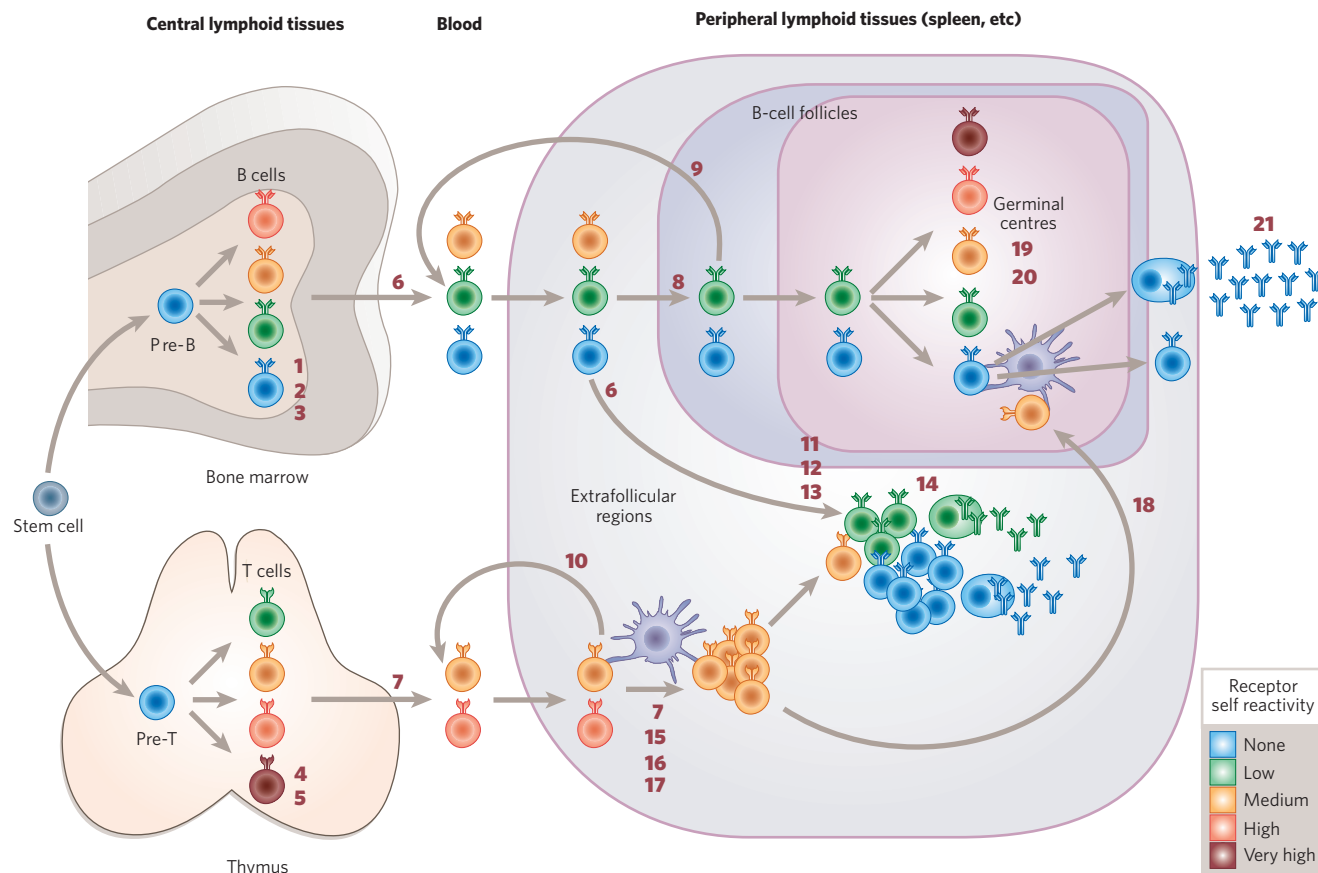


Figure 2 | A map of the cellular checkpoints regulating self-reactive receptors. The range of different BCRs or TCRs, with varying degrees of avidity for self antigens, are denoted by different colours as marked in the key. Specific cellular mechanisms are shown by red numbers. BCR tolerance mechanisms in central lymphoid organs: (1) Arrest of immature B-cell maturation; (2) BCR light chain editing by V(D)J recombination; (3) Death and deletion of immature B cells. TCR tolerance mechanisms in central lymphoid organs: (4) TCR α -chain editing by V(D)J recombination; (5) Death and deletion of semi-mature T cells. Intrinsic regulation of self-reactive receptors by anergy and biochemical tuning: (6) BCR tuning/anergy; (7) TCR tuning/anergy. Extrinsic regulation of self-reactive receptors by competitive mechanisms: (8) Follicular exclusion of B cells; (9) B-cell competition for

BAFF; (10) T-cell competition for IL-7.

Extrinsic regulation of self-reactive receptors by limiting immunogenic co-stimuli: (11) Controls on availability of extrafollicular T-cell help; (12) Control of TLR ligands and signalling; (13) B-cell death induced by FASL from T cells; (14) BCR inhibition of plasma-cell differentiation; (15) Control of B7 ligands and other costimulatory molecules; (16) T-cell death induced by FASL; (17) T-cell suppression by T_R cells.

Regulation of self-reactive receptors in follicles: (18) Control of ICOS and follicular T helper cell differentiation; (19) BCR-induced death of germinal-centre B cells; (20) Germinal-centre B-cell death from competition for follicular T helper cells.

Tolerance of self-reactive receptors at the final effector phase: (21) Control of autoantibody accumulation and inflammation in tissues.

nation bind with a potentially dangerous affinity to a self antigen¹⁻⁴. Since only 3–8% of the population develops an autoimmune disease⁵, it is remarkable that this enormous burden of self-reactive receptors is so well regulated in most of us.

Each lymphocyte usually produces only a single receptor out of the billions possible. Experiments have established that if this receptor is self reactive, then four cellular strategies are employed to deal with them (Fig. 1). First, the cell displaying the 'forbidden', or self-reactive, receptor can be triggered to die, as originally envisaged in Burnet's concept of clonal deletion. Second, a cell bearing a forbidden receptor can 'edit' the offending receptor by further V(D)J recombination or somatic hypermutation to display a different receptor that is not self reactive⁶. Third, intrinsic biochemical and gene-expression changes can reduce the ability of the cell to be triggered by self-reactive receptors. This is generally termed clonal anergy or tuning⁷⁻⁹. Finally, even if the cells have evaded the three mechanisms above, collectively called 'immunological ignorance', extrinsic controls can limit the danger of self-reactive receptors. These extrinsic controls limit the supply of essential growth factors, costimuli, pro-inflammatory mediators and other factors, and also include active suppression by regulatory T (T_R) cells, through a mechanism that is poorly understood. The

latter topic is reviewed separately in this issue by Kronenberg and Rudensky (page 598).

These four mechanisms act as checkpoints on the pathway leading to the production of secreted antibodies and effector T cells. The cellular map in Fig. 2 derives primarily from experiments tracing the fate of forbidden receptors and the cells bearing them in mice that had exceptionally high frequencies of receptor for particular antigens. In humans, cells with self-reactive receptors are too heterogeneous and infrequent to visualize these cellular mechanisms at work. The gulf between experimental animals and the clinic is now being bridged by the discovery of individual genes and proteins that are essential both for specific cellular tolerance mechanisms in mice and for preventing autoimmunity in humans. Here, we review the state of the field, following the map shown in Fig. 2. We focus on how our understanding of cellular and gene maps of self tolerance and immunity will guide rational solutions to autoimmune disease.

BCR tolerance mechanisms in central lymphoid organs

Analysis of transgenic mice with a restricted repertoire of BCRs revealed a series of cellular events that are triggered immediately after an immature B cell displays a self-reactive receptor in the bone

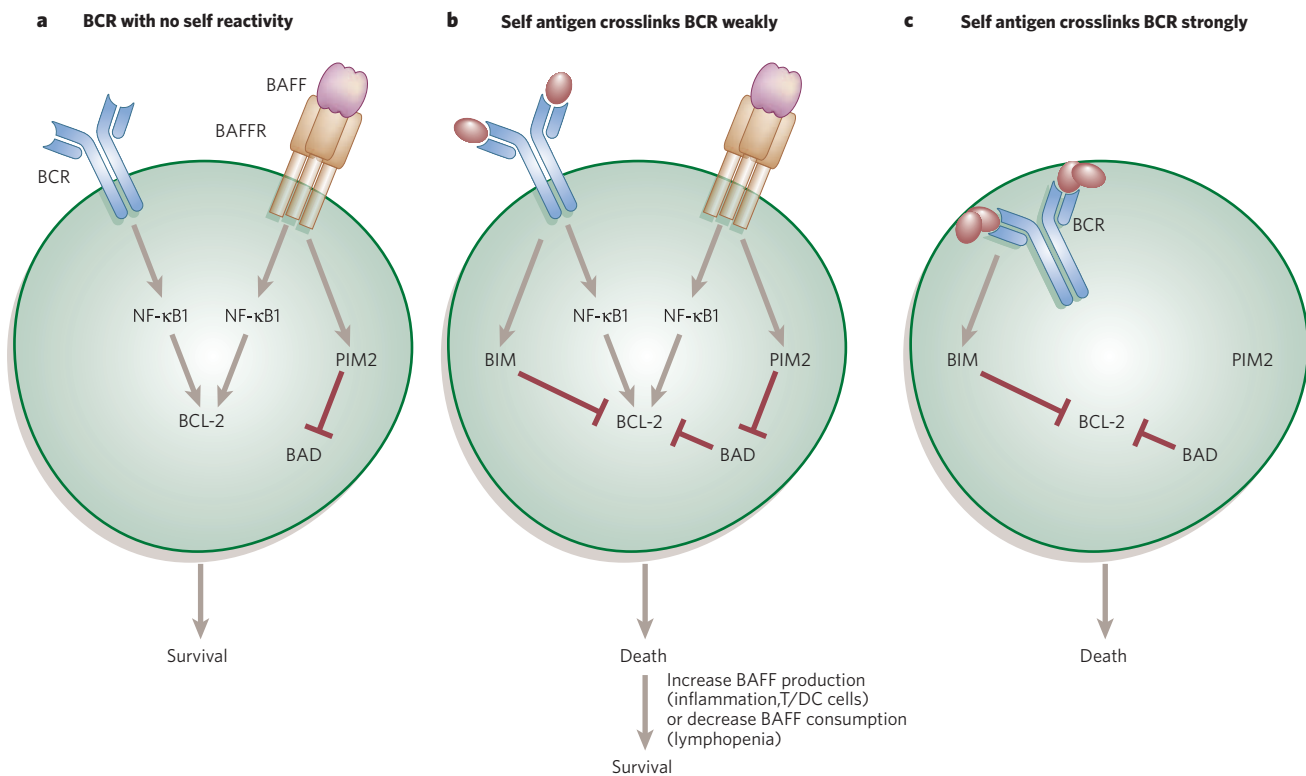


Figure 3 | Integration of intrinsic and extrinsic apoptotic controls deletes cells with self-reactive BCRs. Three examples are shown of what is likely to be a continuum. **a**, A BCR with no self-reactivity, where BCR and BAFFR survival pathways dominate. **b**, A BCR with intermediate self-reactivity,

where BCR signalling activates survival and death pathways, so that death dominates unless increased BAFF is supplied. **c**, A BCR with avid self-reactivity, where BCR internalization and maturation arrest cripples the BCR and BAFFR survival pathways, while BIM induction promotes death.

marrow. If the strength of receptor crosslinking and intracellular signalling exceeds a certain threshold, the immature B cell rapidly internalizes the offending BCR and temporarily halts its maturation programme^{6,10,11}. This has three consequences. First, homing receptors, such as CD62 ligand (CD62L), are not expressed. Such receptors are needed for B cells to enter the lymph nodes¹⁰. Second, receptors for B-cell-activating factor (BAFF), a circulating cytokine required to sustain peripheral B-cell survival, are poorly induced¹². Third, *RAG1* (recombination-activating gene 1) and *RAG2*, which encode the core enzymes for V(D)J recombination, continue to be expressed. This allows the BCRs to be edited by rearranging a replacement BCR light chain^{6,13}.

If a B cell with a forbidden receptor fails to edit to a less self-reactive receptor, cell death occurs within 1–2 days, either in the bone marrow or shortly after arriving in the spleen¹⁰. The process of clonal deletion may result partly from growth-factor withdrawal, owing to low expression of BAFF receptors on immature B cells. Deletion also involves BCR-induced cell death through increasing the levels of BIM (BCL-2-interacting mediator of cell death). This pro-apoptotic factor inhibits essential B-cell survival proteins from the BCL-2 family¹⁴ (Fig. 3c). Interestingly, BIM-deficient mice spontaneously produce anti-DNA autoantibodies after a latent period of many months¹⁴.

It is not known whether defects in BCR editing or deletion contribute to human autoimmunity. Antibodies that recognize human nuclear antigens and DNA — such as those found in patients with systemic lupus erythematosus (SLE) — are more frequently borne by the subset of immature B cells with little or no surface BCR in human bone marrow⁴. This suggests that the BCRs have been internalized and provides evidence for conservation in humans of the processes described above. As opposed to B cells producing antibodies to systemic antigens, which can be regulated in this way, B cells bearing autoantibodies to organ-specific antigens¹⁵ such as those causing Graves' disease are not encountered in the bone marrow. The failure to edit or delete

this important class of self-reactive BCRs may put extra pressure on other mechanisms described below.

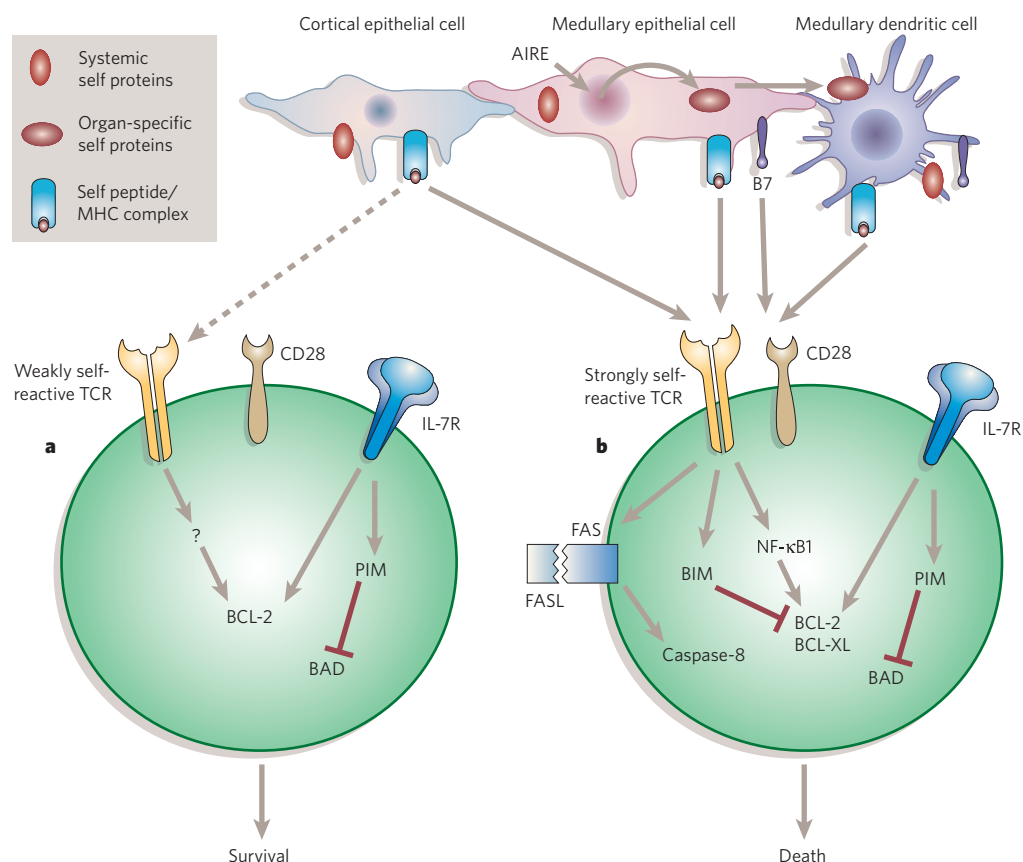
TCR tolerance mechanisms in central lymphoid organs

Receptor editing and cellular deletion also accompany V(D)J recombination in developing T cells in the thymus⁶, although deletion appears to be the predominant process. Unlike BCRs, which are designed to recognize native antigen, TCRs are selected to recognize a composite ligand comprising peptide fragments of antigen bound to MHC molecules. Composites of self peptides and MHC are displayed on the surface of cortical thymic epithelial cells, and TCRs that weakly bind these ligands trigger maturation signals that inhibit RAG gene expression (thereby closing off the option of editing)¹⁶, increase TCR cell-surface expression and induce the expression of homing receptors for chemokines found in the thymic medulla and the peripheral lymphoid tissues. The thymic cortical epithelium is unique in supporting this essential process of positive selection, through as yet unknown factors³. A minority of self-reactive TCRs trigger an editing process; in this case, TCRs are downregulated, RAG expression continues and the offending TCR α -chain is replaced or diluted with a second α -chain that is less self reactive⁶.

As positively selected thymocytes move from the cortex towards the medulla, they continue to test their TCRs for self reactivity, but now this occurs on medullary thymic epithelial cells and dendritic cells of haemopoietic origin¹⁶ (Fig. 4). These medullary cells express T-cell costimulatory molecules, such as CD80 (also known as B7.1) and CD86 (B7.2), the ligands for CD28. At this point, TCRs that bind strongly to self-peptide–MHC combinations trigger the death (negative selection) of thymocytes. The crucial role of medullary cells in ensuring self tolerance is clearly illustrated in studies of mice where medullary MHC molecules are missing or B7 molecules are missing or blocked by antibodies. In such animals, T cells with self-reactive TCRs reach the periphery and cause systemic inflammatory conditions

Figure 4 | Map of pathways for deleting T cells with strongly and weakly self-reactive TCRs in the thymus. Two examples are shown of what is likely to be a continuum:

a, a TCR with weak self reactivity, where TCR and IL-7R pro-survival pathways dominate; **b**, a TCR with strong self reactivity where TCR-induced BIM and FAS death pathways predominate. Similar pathways also act in mature peripheral T cells, where competition for limiting IL-7 and self-peptide/MHC, and induction of B7 ligands for CD28, add extra levels of extrinsic control.



resembling graft-versus-host disease^{3,17}. The well established association between particular MHC molecules and susceptibility to specific autoimmune diseases may stem from inefficient presentation of particular self peptides during this phase of TCR deletion^{18,19}. Development of peripheral nerve-specific autoimmunity in CD86-deficient non-obese diabetic (NOD) mice²⁰ may arise from a similar problem with thymic deletion of nerve-specific self peptides.

A strong connection between thymic TCR deletion and human autoimmune disease has been forged by understanding the function of the autoimmune regulator (*AIRE*) gene. Crippling mutations in the human gene are responsible for autoimmune polyendocrine syndrome 1 (refs 21, 22), an infrequent but devastating disorder that targets many discrete organs. Corresponding *Aire* mutations in mice cause similar, albeit milder, organ-specific autoimmunity^{23,34} and a marked failure to delete organ-specific TCRs in the thymus because tissue-specific genes are not switched on in rare medullary thymic epithelial cells^{24–26}. Promiscuous expression of many different tissue-specific proteins, such as insulin, was noted in rare medullary thymic cells over a decade ago. This expression was proposed as a way to couple central thymic tolerance mechanisms to peripherally expressed self proteins²⁷. The crucial role of this mechanism is established by the severe autoimmune syndrome in *AIRE*-deficient humans, and it may explain why inherited promoter variants that decrease thymic expression of the insulin gene are associated selectively with autoimmune diabetes in humans^{28,29}.

A less-defined abnormality of medullary thymic epithelium appears to be responsible for myasthenia gravis, where hyperproliferative or neoplastic thymic epithelial cells that display subunits of the acetylcholine receptor activate T cells and precipitate the formation of circulating anti-acetylcholine receptor autoantibodies³⁰. Surgical removal of the thymus in some cases cures the autoimmunity.

The signalling events involved in negative selection of T cells are still poorly understood. The tyrosine kinase ZAP70 (ζ -chain-associ-

ated protein kinase of 70 kDa) is required for thymocyte deletion because partial deficiency of ZAP70 in mice allows cells with forbidden TCRs to escape death and cause a systemic inflammatory disorder resembling rheumatoid arthritis³¹. The onset of negative selection also requires GRB2 (growth-factor-receptor-bound protein 2) (ref. 32) and MINK (misshapen-Nck-interacting kinase (NIK) related kinase) (ref. 33) plus intense but transient activation of ERK (extracellular signal-regulated kinase) and prolonged p38 and JNK (Jun kinase) activation¹⁶. At the distal end of the pathway, deletion partly requires induction of BIM expression at the messenger RNA and protein level¹⁴. BIM antagonizes BCL-2 and related proteins to release pro-apoptotic BAX and BAK, which are also required for deletion³⁴. Deletion also involves induction of members of the Nur77 family of orphan nuclear receptors, and TCR-induced thymocyte death is blocked by a dominant-negative Nur77 mutant³⁵. Induction of BIM and Nur77 expression and cell death in thymocytes with forbidden TCRs is selectively defective in the NOD mouse strain, owing to the cumulative T-cell-intrinsic effects of four of the chromosomal loci that contribute to diabetes susceptibility in this strain³⁶. The resistance of NOD thymocytes to deletion, observed in three separate experimental systems^{37–39}, thus appears to be an important component of the susceptibility of this strain to a range of autoimmune diseases.

The combination of strong stimulatory signals through TCR and CD28 is, paradoxically, a potent trigger of nuclear factor- κ B (NF- κ B) activation, the pro-survival pathway that induces expression of BCL-2 proteins in mature peripheral T cells⁴⁰. NF- κ B activation may counteract BIM-induced cell death (Fig. 4), and indeed gene-expression profiling demonstrates specific induction of NF- κ B genes in semimature thymocytes with self-reactive TCRs³⁶. Strong TCR engagement also induces expression of an extracellular protein, FAS ligand (FASL, also known as CD95L), and triggers T-cell death independently from the BIM/BCL-2 mechanism through FASL–FAS interaction and the caspase-8 proteolytic cascade. The requirement for this

pathway in thymocyte deletion, though, is still debated^{41,42}. The severe systemic autoimmune syndromes that develop in mice lacking BIM¹⁴, and in both mice and humans with defects in FASL and FAS⁴³, underscore the crucial roles of BIM and FAS at several T- and B-cell tolerance checkpoints.

Marked disruption of thymic deletion by defects in a single gene, exemplified by homozygous *AIRE* deficiency, is an informative but apparently rare cause of human autoimmune disease. More commonly, autoimmune disease reflects subtle decreases at multiple points in the thymic deletion pathways (summarized in Fig. 4), as seen in the NOD mouse³⁶. It is striking that loss of one copy of a key gene, such as *Aire*²⁶ and *Bim*¹⁴, in this pathway is sufficient to create a small but measurable decrease in the efficiency of deletion. Equally small changes in insulin gene expression in the thymus correlate with human susceptibility to type 1 diabetes^{28,29}. These findings indicate that the process of thymic deletion is on a knife-edge rather than buffered by a large safety margin for deleting forbidden receptors. Therefore, there is likely to be large individual variation in the efficiency of deleting particular TCRs among people.

Intrinsic regulation by anergy and biochemical tuning

In addition to deletion and editing, self-reactive receptors are regulated in primary and secondary lymphoid tissues by intrinsic biochemical changes in the cells displaying them. Several intrinsic cellular mechanisms of anergy are well documented in B cells with self-reactive BCRs (refs 7, 11, 44; Fig. 1c). The first is decreased display of self-reactive BCRs on the cell surface, varying from as little as 50% to more than 99% reduction, owing to accelerated endocytosis (M. Blery, unpublished observations) and blocked transport of new BCRs out of the endoplasmic reticulum⁴⁵. Partly independent from receptor downregulation, self-reactive BCRs activate tyrosine kinase signalling poorly, limiting cell survival because of weak NF- κ B1 activation. The self-reactive BCRs nevertheless continue to signal for BIM induction to promote death⁴⁶ and to ERK pathways that block Toll-like receptor 9 (TLR9)-induced differentiation into plasma cells⁴⁷. Each of these changes is reversible if the BCR stops binding to the self antigen, as would occur if the BCR is edited to lose self reactivity through somatic hypermutation in germinal centres.

Two other mechanisms of molecular feedback dampen self-reactive BCR signalling. Purely biochemical tuning involves proteins that increase the threshold for B-cell activation and are expressed constitutively regardless of BCR specificity. An example is the recruitment of the tyrosine phosphatase SHP1 (SH2-domain-containing protein tyrosine phosphatase 1) to the activated BCR through the cell-surface proteins CD22 and PD1. Another example is the recruitment of the lipid phosphatase SHIP (SH2-domain-containing inositol-5-phosphatase) to the activated BCR through Fc receptor- γ (refs 7, 48). Defects in either of these pathways dysregulate responses to foreign and self antigens, and create susceptibility to spontaneous autoantibody production. The changes in gene expression that occur selectively in cells bearing self-reactive BCRs are a second adaptation. An example is expression of the CD5 cell-surface protein induced selectively by self-reactive BCRs, which provides an additional inhibitory receptor to recruit SHP1 and inhibit BCR signalling and activation⁴⁹.

Biochemical and genetic feedback loops are equally important for tuning T cells because all the T cells that leave the thymus will have been selected to have a moderate degree of TCR self reactivity^{8,9,16}. Whereas antigen-receptor downregulation is generally less marked in self-reactive T cells than B cells, expression of the inhibitory receptor CD5 is induced to 10–50-fold higher levels in self-reactive T cells than in B1 cells or anergic B cells. CD5 levels are dynamically adjusted on individual thymic and peripheral T cells in proportion to the strength of TCR self reactivity, downregulating the response of TCRs to self peptides to avoid T-cell activation or deletion^{50,51}. Expression of another inhibitory receptor, cytotoxic T-lymphocyte antigen 4 (CTLA4), is induced at a high threshold of TCR self reactivity and inhibits T-cell activation by competing with CD28 for ligation with B7

molecules and by transmitting inhibitory signals^{48,52,53}. Lack of CTLA4 causes massive accumulation of self-reactive T cells in peripheral lymphoid and nonlymphoid tissues, both by disrupting intrinsic regulation of TCR-induced proliferation^{53,54} and by impairing the suppressive function of T_R cells. Subtle functional variants of the *CTLA4* gene are associated with susceptibility to thyroid autoimmunity and type 1 diabetes in humans and mice⁵⁵.

Increased expression of the ubiquitin ligases CBL-B, GRAIL and ITCH can also accompany chronic TCR signalling *in vitro*^{56–58}. These proteins interfere with TCR, CD28 and cytokine receptor signalling by tagging the TCR–CD28 or cytokine receptor signalling molecules with ubiquitin. This tag can trigger endocytosis and alter intracellular trafficking of TCRs, promote proteolytic degradation of receptors or signalling subunits, or allosterically interfere with signalling^{56,57,59,60}. It is not known whether these ubiquitin ligases are induced to higher levels in strongly self-reactive T cells *in vivo*. But their importance for preventing autoimmunity in rodents is clear: *cbl-b* deficiency coupled with a particular MHC haplotype causes type 1 diabetes in the KDP (Komeda diabetes prone) rat strain⁶¹. Mice lacking *Itch*, or *cbl-b* and its close relative *c-cbl*, develop large numbers of activated T cells and produce high titres of autoantibodies^{59,60}. Although the animal studies indicate crucial roles for anergy and tuning, little is known about the genes involved in human autoimmune disease other than the association with CTLA4 variants discussed above.

Extrinsic regulation of self-reactive receptors

A significant fraction of self-reactive BCRs and TCRs fail to be edited or trigger deletion in primary lymphoid tissues, either because the self antigens are bound with only low avidity or because they are not sufficiently abundant in the primary lymphoid organs. For receptors with intermediate avidity for self antigens, the risk they pose for autoimmunity may not overshadow their potential use in fighting infection. B cells with receptors that fall into this zone undergo a conditional type of clonal deletion that is extrinsically regulated through competition with B cells bearing less self-reactive BCRs (ref. 62). The recently revealed molecular basis^{46,63} for this illuminates general principles of broad significance (Fig. 3b).

The survival of peripheral B cells depends on BAFF, which is produced in limiting quantities primarily by radioresistant lymphoid stromal cells¹². BAFF activates its receptor, BAFFR, on B cells and triggers an increase in the activity of the transcription factor NF- κ B2, which maintains peripheral B-cell survival through induction of BCL-2 expression^{12,64}. BAFF also induces expression of the serine–threonine kinase, PIM2 (refs 46, 65), which has potent pro-survival effects by phosphorylating and inhibiting the activity of the pro-apoptotic protein BAD⁶⁶. Constant engagement of self-reactive BCRs, below the threshold required to trigger maturation-arrest in the bone marrow, is still sufficient to increase BIM expression and consequently elevate the survival requirement for BAFF (ref. 46; Fig. 3b). With large numbers of circulating B cells, the self-reactive cells fail to receive enough BAFF and are competitively deleted. This mechanism provides a 'survival of the fittest' process that can select against subtle differences in affinity for self antigen⁶³. Conversely, self-reactive BCRs are more likely to survive when competition for BAFF is reduced, for example, in states of B-cell lymphopenia, or when BAFF synthesis is increased during infection or in pathological conditions^{12,46,63}. In a clinical setting, partial antagonism of BAFF may be a particularly powerful way to enhance autoimmune B-cell sensitivity to natural tolerance mechanisms. It may also be a valuable adjunct to B-cell-depleting therapeutics, such as anti-CD20 (Rituxan), which might otherwise increase BAFF availability.

In common with B cells, the survival of mature T cells depends upon continuous signalling in the peripheral lymphoid tissues. Keeping T cells alive requires TCR signalling through contact with ubiquitous MHC ligands as well as exposure to interleukin-7 (IL-7) (refs 67–69). Normally, IL-7 levels are low and maintain T cells in interphase, and strongly self-reactive TCRs trigger only a transient proliferation that is

overwhelmed by BIM-induced cell death¹⁴. In the case of T lymphocytopenia, however, IL-7 levels rise and amplify TCR signalling, causing naive T cells to proliferate and fill up 'space'. Such homeostatic proliferation may activate T cells reactive to tissue-specific antigens and promote migration of these cells into extralymphoid sites, thereby risking the development of autoimmune disease at these sites. This scenario may explain why T lymphocytopenia predisposes people to autoimmune disease. Thus, some children with limited T- and B-cell production owing to a partial loss of RAG1 or RAG2 V(D)J recombinase activity develop Omenn syndrome, a massive expansion of peripheral T cells that infiltrate the skin and intestine, which resembles graft-versus-host disease⁷⁰. In Wiskott–Aldrich syndrome, lymphopenia and defective T-cell function are accompanied by an array of autoimmune and inflammatory conditions⁷¹. Relaxation of competitive regulation may also explain the development of thyroid autoimmunity in one-third of patients with multiple sclerosis treated with the lymphocyte-depleting antibody, CAMPATH-1H (ref. 72). T-cell lymphopenia is an essential contributing factor to autoimmune diabetes susceptibility in the BB rat strain^{73,74} and is implicated in causing autoimmunity in the NOD mouse⁷⁵.

Extrinsic regulation by limiting immunogenic costimuli

Antibody responses specific for many antigens depend upon B cells receiving two signals in short succession: signal one from the antigen binding to the BCR, and signal two from T helper cells. The T-cell surface protein CD40 ligand (CD40L) and secreted cytokines IL-2, IL-4, IL-5 and IL-21 comprise the second signals for B-cell proliferation and differentiation into antibody-secreting plasma cells^{76,77}. Owing to thymic tolerance, the capacity of T cells to provide help for self-reactive B cells is limited. Nevertheless, signal two from T helper cells responding to foreign antigen can be misdirected to self-reactive BCRs that cross-react with a component of a microorganism or recognize a self component that associates with microbial components. It can also be misdirected by delivery of helper signals to bystander B cells. However, even in the rare instances when infection can be shown to trigger autoantibody diseases, notably Guillain–Barré syndrome, where cross-reactivity between antigens from *Campylobacter jejuni* and components of peripheral nerves elicits neuropathic autoantibodies⁷⁸, only one in 1,000 infected individuals develop the autoantibodies. The resistance of 99.9% of people must be accounted for by efficient B-cell-intrinsic tolerance mechanisms.

A second pathway for antibody production involves the signalling of B cells through contact with stimulatory ligands released by microorganisms, notably bacterial flagellins, cell-wall lipopolysaccharides and unmethylated CpG dinucleotides. These are recognized by a set of TLRs on B cells⁷⁹, and can partially substitute for T-cell help to allow antibody control of many microorganisms even in T-cell-deficient individuals. Relatively little is known about how TLR signalling is dampened to ensure that the threshold for stimulation is high enough to stop activation of self-reactive B cells. Experiments in mice have shown that dysregulated activity of the TLR9 pathway (which senses CpG DNA), owing to pathological accumulation of circulating IgG–self-DNA complexes, is a potent driver of the production of autoantibodies against IgG and DNA (ref. 80). TLR9 signalling may explain the production of anti-DNA autoantibodies in a subset of people given procainamide (a drug used to correct an irregular heart-beat), through interference with DNA methyltransferases that mask CpG motifs in our own DNA (ref. 81). It may also explain why SLE can sometimes be treated with chloroquine, an inhibitor of TLR9 signalling. Inadequate clearance of apoptotic cells with exposed CpG DNA and other nuclear antigens may account for the striking association between SLE and genetic deficiencies in classical complement pathway components⁸².

Activation of mature T cells requires a combination of TCR ligation plus a costimulatory signal. T-cell tolerance is favoured when antigen is recognized without immunogenic costimuli. Although there are multiple T-cell costimuli, interaction of CD28 on T cells with the B7

proteins CD80 and CD86 on antigen-presenting cells is especially important⁵². Expression of B7 proteins and other costimulatory ligands is induced on the surface of B cells and dendritic cells in response to TLR signalling and potentiates the survival and clonal expansion of T cells, in part by activating the NF- κ B1 pathway and inducing the expression of BCL-2-related pro-survival proteins⁴⁰. When B7–CD28 signalling is blocked, TCR-induced proliferation is cut short and T-cell death ensues, as a result of either BIM expression or FAS–FASL signalling^{52,53,68}. Blocking B7–CD28 costimulation is thus a promising avenue to switch immune responses into tolerance, for example by using a competing soluble receptor for B7 proteins such as a CTLA4–immunoglobulin fusion protein (see the review in this issue by Feldmann and Steinman, page 612). This treatment may also interfere with tolerance by decreasing thymic deletion, T_R function (see the review by Kronenberg and Rudensky, page 598) and intrinsic T-cell regulation by CTLA4.

Regulation of self-reactive receptors in follicles

Each of the checkpoints described above deals with self-reactive receptors generated by V(D)J recombination in the primary lymphoid organs; however, self-reactive BCRs are also generated in a second wave of receptor-gene diversification through somatic hypermutation in germinal-centre follicles of peripheral lymphoid tissues^{30,83,84} (Fig. 2). Somatic hypermutation poses a particularly severe threat of autoimmunity for three reasons. First, this process can markedly increase the affinity of antibodies for self antigens, so that the same concentration of secreted antibody may cause 100-fold more damage. Second, the follicular pathway of B-cell differentiation generates long-lived plasma cells and memory cells, which can continue to produce antibody indefinitely⁸⁵. Third, autologous DNA, an important self-antigen target in SLE, is presented by the numerous apoptotic cells⁸⁶ in germinal centres, and organ-specific self components are trapped and displayed as immune complexes on follicular dendritic cells in autoimmune disease^{30,87}, where they represent a powerful potential stimulus for autoantibody production.

In animal models of SLE, most, if not all, anti-double-stranded DNA antibodies are somatically mutated, and the pattern of somatic mutations indicates that autoreactive B cells are positively selected for higher affinity⁸³. Germinal centres and hypermutated, autoantigen-specific BCRs form in ectopic sites immediately adjacent to sources of autoantigen in a range of human autoimmune diseases^{30,87}, as well as arising in lymph nodes that drain organs affected by autoimmunity in experimental animals⁸⁸.

Remarkably little is known about the mechanisms that normally deal with self-reactive BCRs arising in germinal centres or how these are disrupted in human autoimmunity. Self-reactive BCRs induce subtle differences in B-cell responsiveness to the opposing chemokine gradients between follicles and extrafollicular zones^{11,62,89}, excluding them from follicles and thus minimizing their participation in germinal-centre responses. Rapid deletion of B cells within germinal centres can be induced in less than 4 hours — comparable to or faster than the rapid death of self-reactive thymocytes — when the BCR binds to an antigen that is not recognized by the T cells in the germinal centre^{90,91}. In common with deletion of thymocyte and extrafollicular B cells, deletion of germinal-centre B cells may involve the induction of BIM expression or trigger the FAS death receptor. Like extrafollicular B cells (Fig. 2), self-reactive B cells in the germinal centres may be extrinsically regulated by competition for BAFF, or by competition for CD40L, IL-21 and ICOS (inducible T-cell costimulator) signals from follicular T cells (see below). Cells with cross-reactive BCRs that recognize both self and foreign antigen⁸⁴ would, in principle, be able to survive, but self-antigen-induced BIM expression would put them at a disadvantage when competing with cells bearing purely foreign reactive BCRs whose BCR engagement is less chronic and more tightly linked to T-cell help.

The small subset of CD4⁺ T cells found in germinal centres follows a unique programme of differentiation compared with their

extrafollicular counterparts. Like extrafollicular helper T cells, follicular T cells express CD40L, which is required to sustain survival and proliferation of germinal-centre B cells⁹². Follicular T cells, however, also display high levels of ICOS, which is specifically required to help germinal-centre antibody responses in mice and humans^{52,93}. Follicular T cells also require costimulation through OX40L (ref. 94) and an intracellular signalling adaptor for the SLAM family of costimulatory receptors, SH2D1a (also known as SAP) (ref. 95). The fact that antigen stimulation in the absence of microbial TLR agonists, as would normally occur for self antigens, does not induce T-cell entry into follicles⁹⁶ suggests that strict regulation of follicular T-helper cell differentiation may block self-reactive T cells from delivering help to germinal-centre B cells. Recently, we discovered a novel ubiquitin ligase, Roquin, that is essential for repressing ICOS expression in T cells. When Roquin is defective in mice, self-antigen-directed follicular T-cell help becomes uncontrolled, generating large numbers of germinal centres and extraordinary levels of autoantibody production⁹⁷. These advances begin to open up analysis of the tolerance checkpoints controlling autoantibody selection during the crucial germinal-centre phase of antibody evolution.

Self-reactive receptor tolerance at the final effector phase

Even when high-affinity antibodies do form and circulate in sufficient quantity to cause disease upon transplacental transfer to a foetus, there is often little or no disease in individuals with these antibodies⁹⁸. Moreover, when pathology does occur, it is often limited to focal points, illustrated by the circumscribed skin lesions in pemphigus despite ubiquitous skin distribution of the target autoantigen. Analysis of a mouse model of rheumatoid arthritis shows that even when sufficient autoantibody is present in the circulation, its capacity to localize in joints to produce joint pathology depends on inflammatory cascades involving Fc receptors, mast cells, neutrophils and complement^{99,100}. There is clearly considerable scope for tolerating self-reactive receptors even at this level, and much more research is needed into this phase of regulation.

Concluding remarks

Although many self-tolerance mechanisms exist, defects in a single checkpoint, such as AIRE, can lead to autoimmune disease. The clinical manifestation is nevertheless seen only after a latent period of many years and then only against a few proteins or organs. There are obvious parallels here with inherited defects in tumour-suppressor genes, favouring the view that successive and parallel tolerance checkpoints provide back-up mechanisms to control all but a few exceptional forbidden receptors. There seems to be hundreds of genes such as *AIRE*, *BIM*, *ZAP70*, *CBLB*, *FAS* and *ROQUIN* involved in these checkpoints, and most of the candidate genes are yet to be discovered on the basis of the rate at which new autoimmunity genes are currently being identified. Since heterozygous mutation in any one of these genes may predispose a person to autoimmunity, the sheer number of genes involved may collectively account for the ~5% of people with autoimmune disease despite a low population frequency for heterozygous mutations in any one of these genes. Population-wide scans based on common DNA polymorphisms will not be effective tools to identify predisposing defects of this type: instead, exon resequencing of individuals with autoimmune disease will be required.

Regardless of whether predisposition to autoimmunity is caused by rare or common genetic variants, interventions aimed at preventing or treating autoimmunity will need to be tailored to correct weak cellular checkpoints, shore up back-up mechanisms and avoid doing more harm by interfering with these mechanisms and thus exacerbating the breakthrough of forbidden receptors. The development of thyroid autoimmunity through lymphopenia as a result of antibody therapy in multiple sclerosis patients⁷² and the development of systemic autoimmunity when B7 molecules are blocked in experimental animals¹⁷ highlight the risks. Well-targeted interventions require a more complete map of the cellular mechanisms and genes underpin-

ning self tolerance, and more ways to test for individual variation. The conserved genes and proteins now laid out by mammalian genome sequencing, and the relative ease of moving back and forward between human and rodent gene analysis, provide the vehicle for solving both of these challenges. ■

- Ignatowicz, L., Kappler, J. & Marrack, P. The repertoire of T cells shaped by a single MHC/peptide ligand. *Cell* **84**, 521–529 (1996).
- Zerrahn, J., Held, W. & Raulet, D. H. The MHC reactivity of the T cell repertoire prior to positive and negative selection. *Cell* **88**, 627–636 (1997).
- Laufer, T. M., DeKoning, J., Markowitz, J. S., Lo, D. & Glimcher, L. H. Unopposed positive selection and autoreactivity in mice expressing class II MHC only on thymic cortex. *Nature* **383**, 81–85 (1996).
- Wardemann, H. et al. Predominant autoantibody production by early human B cell precursors. *Science* **301**, 1374–1377 (2003).
- Jacobson, D. L., Gange, S. J., Rose, N. R. & Graham, N. M. Epidemiology and estimated population burden of selected autoimmune diseases in the United States. *Clin. Immunol. Immunopathol.* **84**, 223–243 (1997).
- Nemazee, D. & Hogquist, K. A. Antigen receptor selection by editing or downregulation of V(D)J recombination. *Curr. Opin. Immunol.* **15**, 182–189 (2003).
- Healy, J. I. & Goodnow, C. C. Positive versus negative signaling by lymphocyte antigen receptors. *Annu. Rev. Immunol.* **16**, 645–670 (1998).
- Schwartz, R. H. T cell anergy. *Annu. Rev. Immunol.* **21**, 305–334 (2003).
- Grossman, Z. & Paul, W. E. Self-tolerance: context-dependent tuning of T cell antigen recognition. *Semin. Immunol.* **12**, 197–203 (2000).
- Hartley, S. B. et al. Elimination of self-reactive B lymphocytes proceeds in two stages: arrested development and cell death. *Cell* **72**, 325–335 (1993).
- Fields, M. L. & Erikson, J. The regulation of lupus-associated autoantibodies: immunoglobulin transgenic models. *Curr. Opin. Immunol.* **15**, 709–717 (2003).
- Mackay, F., Schneider, P., Rennert, P. & Browning, J. BAFF and APRIL: a tutorial on B cell survival. *Annu. Rev. Immunol.* **21**, 231–264 (2003).
- Jankovic, M., Casellas, R., Yannoutsos, N., Wardemann, H. & Nussenzweig, M. C. RAGs and regulation of autoantibodies. *Annu. Rev. Immunol.* **22**, 485–501 (2004).
- Strasser, A. & Bouillet, P. The control of apoptosis in lymphocyte selection. *Immunol. Rev.* **193**, 82–92 (2003).
- Akcaraju, S., Cnaan, K. & Goodnow, C. C. Self-reactive B cells are not eliminated or inactivated by autoantigen expressed on thyroid epithelial cells. *J. Exp. Med.* **186**, 2005–2012 (1997).
- Palmer, E. Negative selection — clearing out the bad apples from the T-cell repertoire. *Nature Rev. Immunol.* **3**, 383–391 (2003).
- Gao, J. X. et al. Perinatal blockade of B7-1 and B7-2 inhibits clonal deletion of highly pathogenic autoreactive T cells. *J. Exp. Med.* **195**, 959–971 (2002).
- Kanagawa, O., Martin, S. M., Vaupel, B. A., Carrasco-Marin, E. & Unanue, E. R. Autoreactivity of T cells from nonobese diabetic mice: an I-Ag7-dependent reaction. *Proc. Natl Acad. Sci. USA* **95**, 1721–1724 (1998).
- Wicker, L. S., Todd, J. A. & Peterson, L. B. Genetic control of autoimmune diabetes in the NOD mouse. *Annu. Rev. Immunol.* **13**, 179–200 (1995).
- Salomon, B. et al. Development of spontaneous autoimmune peripheral polyneuropathy in B7-2-deficient NOD mice. *J. Exp. Med.* **194**, 677–684 (2001).
- Nagamine, K. et al. Positional cloning of the APECED gene. *Nature Genet.* **17**, 393–398 (1997).
- Aaltonen, J. et al. An autoimmune disease, APECED, caused by mutations in a novel gene featuring two PHD-type zinc-finger domains. *Nature Genet.* **17**, 399–403 (1997).
- Ramsey, C. et al. Aire-deficient mice develop multiple features of APECED phenotype and show altered immune response. *Hum. Mol. Genet.* **11**, 397–409 (2002).
- Anderson, M. S. et al. Projection of an immunological self-shadow within the thymus by the Aire protein. *Science* **298**, 1395–1403 (2002).
- Liston, A., Lesage, S., Wilson, J., Peltonen, L. & Goodnow, C. C. Aire regulates negative selection of organ-specific T cells. *Nature Immunol.* **4**, 350–354 (2003).
- Liston, A. et al. Gene dosage-limiting role of Aire in thymic expression, clonal deletion, and organ-specific autoimmunity. *J. Exp. Med.* **200**, 1015–1026 (2004).
- Hanahan, D. Peripheral-antigen-expressing cells in thymic medulla: factors in self-tolerance and autoimmunity. *Curr. Opin. Immunol.* **10**, 656–662 (1998).
- Pugliese, A. et al. The insulin gene is transcribed in the human thymus and transcription levels correlated with allelic variation at the INS VNTR-IDDM2 susceptibility locus for type 1 diabetes. *Nature Genet.* **15**, 293–297 (1997).
- Vafiadis, P. et al. Insulin expression in human thymus is modulated by INS VNTR alleles at the IDDM2 locus. *Nature Genet.* **15**, 289–292 (1997).
- Shiono, H. et al. Scenarios for autoimmunization of T and B cells in myasthenia gravis. *Ann. NY Acad. Sci.* **998**, 237–256 (2003).
- Sakaguchi, N. et al. Altered thymic T-cell selection due to a mutation of the ZAP-70 gene causes autoimmune arthritis in mice. *Nature* **426**, 454–460 (2003).
- Gong, Q. et al. Disruption of T cell signaling networks and development by Grb2 haploid insufficiency. *Nature Immunol.* **2**, 29–36 (2001).
- McCarty, N. et al. Signaling by the kinase MINK is essential in the negative selection of autoreactive thymocytes. *Nature Immunol.* **6**, 65–72 (2005).
- Rathmell, J. C., Lindsten, T., Zong, W. X., Cinalli, R. M. & Thompson, C. B. Deficiency in Bak and Bax perturbs thymic selection and lymphoid homeostasis. *Nature Immunol.* **3**, 932–939 (2002).
- Zhou, T. et al. Inhibition of Nur77/Nurrl leads to inefficient clonal deletion of self-reactive T cells. *J. Exp. Med.* **183**, 1879–1892 (1996).
- Liston, A. et al. Generalized resistance to thymic deletion in the NOD mouse; a polygenic trait characterized by defective induction of Bim. *Immunity* **21**, 817–830 (2004).
- Kishimoto, H. & Sprent, J. A defect in central tolerance in NOD mice. *Nature Immunol.* **2**, 1025–1031 (2001).
- Lesage, S. et al. Failure to censor forbidden clones of CD4 T cells in autoimmune diabetes. *J. Exp. Med.* **196**, 1175–1188 (2002).
- Choisy-Rossi, C. M., Holl, T. M., Pierce, M. A., Chapman, H. D. & Serreze, D. V. Enhanced

- pathogenicity of diabetogenic T cells escaping a non-MHC gene-controlled near death experience. *J. Immunol.* **173**, 3791–3800 (2004).
40. Kane, L. P., Lin, J. & Weiss, A. It's all Rel-ative: NF- κ B and CD28 costimulation of T-cell activation. *Trends Immunol.* **23**, 413–420 (2002).
 41. Sprent, J. & Kishimoto, H. The thymus and negative selection. *Immunol. Rev.* **185**, 126–135 (2002).
 42. Villunger, A. *et al.* Negative selection of semimature CD4(+)8(-)HSA+ thymocytes requires the BH3-only protein Bim but is independent of death receptor signaling. *Proc. Natl Acad. Sci. USA* **101**, 7052–7057 (2004).
 43. Nagata, S. Human autoimmune lymphoproliferative syndrome, a defect in the apoptosis-inducing Fas receptor: a lesson from the mouse model. *J. Hum. Genet.* **43**, 2–8 (1998).
 44. Benschop, R. J. *et al.* Activation and anergy in bone marrow B cells of a novel immunoglobulin transgenic mouse that is both hapten specific and autoreactive. *Immunity* **14**, 33–43 (2001).
 45. Bell, S. E. & Goodnow, C. C. A selective defect in IgM antigen receptor synthesis and transport causes loss of cell surface IgM expression on tolerant B lymphocytes. *EMBO J.* **13**, 816–826 (1994).
 46. Lesley, R. *et al.* Reduced competitiveness of autoantigen-engaged B cells due to increased dependence on BAFF. *Immunity* **20**, 441–453 (2004).
 47. Rui, L., Vinuesa, C. G., Blasioli, J. & Goodnow, C. C. Resistance to CpG DNA-induced autoimmunity through tolerogenic B cell antigen receptor ERK signaling. *Nature Immunol.* **4**, 594–600 (2003).
 48. Ravetch, J. V. & Lanier, L. L. Immune inhibitory receptors. *Science* **290**, 84–89 (2000).
 49. Hippen, K. L., Tze, L. E. & Behrens, T. W. CD5 maintains tolerance in anergic B cells. *J. Exp. Med.* **191**, 883–890 (2000).
 50. Wong, P., Barton, G. M., Forbush, K. A. & Rudensky, A. Y. Dynamic tuning of T cell reactivity by self-peptide-major histocompatibility complex ligands. *J. Exp. Med.* **193**, 1179–1187 (2001).
 51. Smith, K. *et al.* Sensory adaptation in naive peripheral CD4 T cells. *J. Exp. Med.* **194**, 1253–1261 (2001).
 52. Sharpe, A. H. & Freeman, G. J. The B7-CD28 superfamily. *Nature Rev. Immunol.* **2**, 116–126 (2002).
 53. Walker, L. S. & Abbas, A. K. The enemy within: keeping self-reactive T cells at bay in the periphery. *Nature Rev. Immunol.* **2**, 11–19 (2002).
 54. Inobe, M. & Schwartz, R. H. CTLA-4 engagement acts as a brake on CD4⁺ T cell proliferation and cytokine production but is not required for tuning T cell reactivity in adaptive tolerance. *J. Immunol.* **173**, 7239–7248 (2004).
 55. Ueda, H. *et al.* Association of the T-cell regulatory gene CTLA4 with susceptibility to autoimmune disease. *Nature* **423**, 506–511 (2003).
 56. Heissmeyer, V. *et al.* Calcineurin imposes T cell unresponsiveness through targeted proteolysis of signaling proteins. *Nature Immunol.* **5**, 255–265 (2004).
 57. Anandasabapathy, N. *et al.* GRAIL: an E3 ubiquitin ligase that inhibits cytokine gene transcription is expressed in anergic CD4⁺ T cells. *Immunity* **18**, 535–547 (2003).
 58. Jeon, M. S. *et al.* Essential role of the E3 ubiquitin ligase Cbl-b in T cell anergy induction. *Immunity* **21**, 167–177 (2004).
 59. Naramura, M. *et al.* c-Cbl and Cbl-b regulate T cell responsiveness by promoting ligand-induced TCR down-modulation. *Nature Immunol.* **3**, 1192–1199 (2002).
 60. Liu, Y. C. Ubiquitin ligases and the immune response. *Annu. Rev. Immunol.* **22**, 81–127 (2004).
 61. Yokoi, N. *et al.* *Cblb* is a major susceptibility gene for rat type 1 diabetes mellitus. *Nature Genet.* **31**, 391–394 (2002).
 62. Cyster, J. G., Hartley, S. B. & Goodnow, C. C. Competition for follicular niches excludes self-reactive cells from the recirculating B-cell repertoire. *Nature* **371**, 389–395 (1994).
 63. Thien, M. *et al.* Excess BAFF rescues self-reactive B cells from peripheral deletion and allows them to enter forbidden follicular and marginal zone niches. *Immunity* **20**, 785–798 (2004).
 64. Claudio, E., Brown, K., Park, S., Wang, H. & Siebenlist, U. BAFF-induced NEMO-independent processing of NF- κ B2 in maturing B cells. *Nature Immunol.* **3**, 958–965 (2002).
 65. Xu, L. G., Wu, M., Hu, J., Zhai, Z. & Shu, H. B. Identification of downstream genes up-regulated by the tumor necrosis factor family member TALL-1. *J. Leukoc. Biol.* **72**, 410–416 (2002).
 66. Fox, C. J. *et al.* The serine/threonine kinase Pim-2 is a transcriptionally regulated apoptotic inhibitor. *Genes Dev.* **17**, 1841–1854 (2003).
 67. Sprent, J. & Surh, C. D. T cell memory. *Annu. Rev. Immunol.* **20**, 551–579 (2002).
 68. Marrack, P. & Kappler, J. Control of T cell viability. *Annu. Rev. Immunol.* **22**, 765–787 (2004).
 69. Barthlott, T., Kassiotis, G. & Stockinger, B. T cell regulation as a side effect of homeostasis and competition. *J. Exp. Med.* **197**, 451–460 (2003).
 70. Le Deist, F., Poinisgon, C., Moshous, D., Fischer, A. & de Villartay, J. P. Artemis sheds new light on V(D)J recombination. *Immunol. Rev.* **200**, 142–155 (2004).
 71. Dupuis-Girod, S. *et al.* Autoimmunity in Wiskott-Aldrich syndrome: risk factors, clinical features, and outcome in a single-center cohort of 55 patients. *Pediatrics* **111**, e622–e627 (2003).
 72. Coles, A. J. *et al.* Pulsed monoclonal antibody treatment and autoimmune thyroid disease in multiple sclerosis. *Lancet* **354**, 1691–1695 (1999).
 73. MacMurray, A. J. *et al.* Lymphopenia in the BB rat model of type 1 diabetes is due to a mutation in a novel immune-associated nucleotide (lan)-related gene. *Genome Res.* **12**, 1029–1039 (2002).
 74. Hornum, L., Romer, J. & Markholst, H. The diabetes-prone BB rat carries a frameshift mutation in lan4, a positional candidate of Iddm1. *Diabetes* **51**, 1972–1979 (2002).
 75. King, C., Ilic, A., Koelsch, K. & Sarvetnick, N. Homeostatic expansion of T cells during immune insufficiency generates autoimmunity. *Cell* **117**, 265–277 (2004).
 76. Foy, T. M., Aruffo, A., Bajorath, J., Buhlmann, J. E. & Noelle, R. J. Immune regulation by CD40 and its ligand gp39. *Annu. Rev. Immunol.* **14**, 591–617 (1996).
 77. Kovanen, P. E. & Leonard, W. J. Cytokines and immunodeficiency diseases: critical roles of the gamma(c)-dependent cytokines interleukins 2, 4, 7, 9, 15, and 21, and their signaling pathways. *Immunol. Rev.* **202**, 67–83 (2004).
 78. Ang, C. W., Jacobs, B. C. & Laman, J. D. The Guillain-Barré syndrome: a true case of molecular mimicry. *Trends Immunol.* **25**, 61–66 (2004).
 79. Beutler, B. Inferences, questions and possibilities in Toll-like receptor signalling. *Nature* **430**, 257–263 (2004).
 80. Leadbetter, E. A. *et al.* Chromatin-IgG complexes activate B cells by dual engagement of IgM and Toll-like receptors. *Nature* **416**, 603–607 (2002).
 81. Richardson, B. DNA methylation and autoimmune disease. *Clin. Immunol.* **109**, 72–79 (2003).
 82. Taylor, P. R. *et al.* A hierarchical role for classical pathway complement proteins in the clearance of apoptotic cells in vivo. *J. Exp. Med.* **192**, 359–366 (2000).
 83. Radic, M. Z. & Weigert, M. Genetic and structural evidence for antigen selection of anti-DNA antibodies. *Annu. Rev. Immunol.* **12**, 487–520 (1994).
 84. Ray, S. K., Putterman, C. & Diamond, B. Pathogenic autoantibodies are routinely generated during the response to foreign antigen: a paradigm for autoimmune disease. *Proc. Natl Acad. Sci. USA* **93**, 2019–2024 (1996).
 85. Slifka, M. K., Antia, R., Whitmire, J. K. & Ahmed, R. Humoral immunity due to long-lived plasma cells. *Immunity* **8**, 363–372 (1998).
 86. Rosen, A. & Casciola-Rosen, L. Clearing the way to mechanisms of autoimmunity. *Nature Med.* **7**, 664–665 (2001).
 87. Weyand, C. M., Kurtin, P. J. & Goronzy, J. J. Ectopic lymphoid organogenesis: a fast track for autoimmunity. *Am. J. Pathol.* **159**, 787–793 (2001).
 88. Mandik-Nayak, L., Wipke, B. T., Shih, F. F., Unanue, E. R. & Allen, P. M. Despite ubiquitous autoantigen expression, arthritogenic autoantibody response initiates in the local lymph node. *Proc. Natl Acad. Sci. USA* **99**, 14368–14373 (2002).
 89. Reif, K. *et al.* Balanced responsiveness to chemoattractants from adjacent zones determines B-cell position. *Nature* **416**, 94–99 (2002).
 90. Shokat, K. M. & Goodnow, C. C. Antigen-induced B-cell death and elimination during germinal-centre immune responses. *Nature* **375**, 334–338 (1995).
 91. Pulendran, B., Kannourakis, G., Nouri, S., Smith, K. G. & Nossal, G. J. Soluble antigen can cause enhanced apoptosis of germinal-centre B cells. *Nature* **375**, 331–334 (1995).
 92. Han, S. *et al.* Cellular interaction in germinal centers: roles of CD40 ligand and B7-2 in established germinal centers. *J. Immunol.* **155**, 556–567 (1995).
 93. Krocze, R. A., Mages, H. W. & Hutloff, A. Emerging paradigms of T-cell co-stimulation. *Curr. Opin. Immunol.* **16**, 321–327 (2004).
 94. Walker, L. S., Gulbranson-Judge, A., Flynn, S., Brocker, T. & Lane, P. J. Co-stimulation and selection for T-cell help for germinal centres: the role of CD28 and OX40. *Immunol. Today* **21**, 333–337 (2000).
 95. Crotty, S., Kersh, E. N., Cannons, J., Schwartzberg, P. L. & Ahmed, R. SAP is required for generating long-term humoral immunity. *Nature* **421**, 282–287 (2003).
 96. Kearney, E. R., Pape, K. A., Loh, D. Y. & Jenkins, M. K. Visualization of peptide-specific T cell immunity and peripheral tolerance induction in vivo. *Immunity* **1**, 327–339 (1994).
 97. Vinuesa, C. G. *et al.* A novel RING-type ubiquitin ligase family member essential to repress follicular helper T cells and autoimmunity. *Nature* doi:10.1038/nature03555 (this issue).
 98. Scofield, R. H. Autoantibodies as predictors of disease. *Lancet* **363**, 1544–1546 (2004).
 99. Wipke, B. T., Wang, Z., Nagengast, W., Reichert, D. E. & Allen, P. M. Staging the initiation of autoantibody-induced arthritis: a critical role for immune complexes. *J. Immunol.* **172**, 7694–7702 (2004).
 100. Monach, P. A., Benoist, C. & Mathis, D. The role of antibodies in mouse models of rheumatoid arthritis, and relevance to human disease. *Adv. Immunol.* **82**, 217–248 (2004).

Acknowledgements This is a broad field, and with very limited space we needed to cite selected reviews and articles. We sincerely apologise for not directly citing all of the important work on which the points discussed are based. We thank our colleagues at The ANU, Oxford University, UCSF, Centenary Institute, Garvan Institute and The Scripps Research Institute for helpful discussions, and thank the Wellcome Trust, NHMRC, JDRF and NIH for grant support.

Competing interests statement The authors declare that they have no competing financial interests.

Regulation of immunity by self-reactive T cells

Mitchell Kronenberg¹ and Alexander Rudensky²

A basic principle of immunology is that lymphocytes respond to foreign antigens but tolerate self tissues. For developing T cells, the ability to distinguish self from non-self is acquired in the thymus, where the majority of self-reactive cells are eliminated. Recently, however, it has become apparent that some self-reactive T cells avoid being destroyed and instead differentiate into specialized regulatory cells. This appears to be beneficial. Subpopulations of self-reactive T cells have a strong influence on self tolerance and may represent targets for therapeutic intervention to control a variety of autoimmune diseases, tumour growth and infection.

The immune system has evolved to recognize and combat infectious agents¹. From a teleological point of view, an ideal set of immune receptors would recognize pathogens but ignore the proteins, DNA and other components that make up our own bodies. Self-reactive cells pose an immediate threat of autoimmunity. In higher organisms, multiple mechanisms of immunological tolerance eliminate or inactivate lymphocytes that bear receptors specific for autoantigens (see the review in this issue by Goodnow *et al.*, page 590). Nevertheless, some autoreactive lymphocyte clones escape these mechanisms and are present within the peripheral lymphocyte pool. One mechanism by which the pathogenic potential of these autoreactive clones is kept in check is through a dedicated lineage of regulatory T (T_R) cells.

The coexistence of autoreactive and protective T cells was revealed by the multi-organ autoimmunity observed in lymphopenic (immune-deficient) recipient mice upon adoptive transfer of naive $CD4^+$ T cells, and by the protection from autoimmune pathology upon co-transfer of a subset of $CD4^+$ T cells expressing interleukin (IL)-2 receptor α -chain (CD25) (ref. 2). Current evidence suggests that the $CD25^+CD4^+$ T cells are themselves self reactive (Fig. 1), and that this property plays an essential role in the commitment to a T_R -cell lineage. Thus, self reactivity can be beneficial as part of a dedicated cellular mechanism preventing autoimmunity.

In addition to $CD25^+CD4^+$ T_R cells, other important self-reactive T-cell sublineages have been identified. Prominent among these are cells that express a semi-invariant T-cell receptor (TCR) specific for conserved self ligands (Fig. 1). These ligands, which are normally present at a low level, might be induced and serve as molecular signs of stress or infection. The best-characterized such T-cell sublineage is the CD1d-dependent natural killer T (NKT) cell. Mucosal invariant T cells, which are reactive with the major histocompatibility complex (MHC) class-I-like molecule MR-1, are a second example of this type of T cell^{3,4}. In contrast to $CD25^+CD4^+$ T_R cells, which have a dedicated suppressor function, NKT cells in some situations facilitate autoimmune pathology, but in others they are part of a protective mechanism.

In this review, we discuss the origin and biology of two distinct lineages of naturally arising self-reactive T cells: $CD25^+CD4^+$ T_R cells and CD1d-dependent NKT cells. Recent evidence suggests that during thymic differentiation, beneficial self reactivity instructs the develop-

ment of specialized populations of T cells. The properties of these different T-cell sublineages, described below, are summarized in Table 1. Here we highlight recent advances including elucidation of the role of the transcription factor Foxp3 as a dedicated T_R -cell lineage specification factor and the identification of natural self and bacterial glycolipids that are recognized by NKT cells.

The T_R -cell lineage

The widespread acceptance of the existence of a dominant tolerance-inducing mechanism that can suppress the response of other immune cells should not be taken for granted. The idea of a suppressive T-cell lineage was out of favour for many years until results from a few key experiments overcame resistance to this notion. This was thanks largely to a functional *in vivo* screen using adoptive cell transfers and anti-CD25 antibody-mediated *in vivo* depletion experiments^{2,5,6}.

Table 1 | Comparison of T_R and iNKT cells

	T_R cells	iNKT cells
TCRs	Diverse	Invariant $V\alpha$
Co-receptors	Mostly CD4	CD4 or double negative (in mice)
Unique genetic requirements	Foxp3, partial for CD28, IL-2	Many, including IL-15 pathway, transcription factors, SAP and genes involved in CD1d antigen presentation
Specificity	Diverse autologous peptides bound to MHC class II Capable of recognizing non-self peptides	Autologous and bacterial glycosphingolipids presented by CD1d
Distribution	Thymus, lymph nodes, spleen, circulation, sites of inflammation	Thymus, spleen, circulation, liver, bone marrow, sites of inflammation
Effector functions	IL-10, TGF- β , CTLA4, cytotoxicity? other?	Diverse T_H1 and T_H2 cytokines
Regulation of autoimmunity	Required to prevent a plethora of autoimmune conditions	Activation of T_H1 and T_H2 responses may prevent or exacerbate autoimmunity

¹La Jolla Institute for Allergy and Immunology, 10355 Science Center Drive, San Diego, California 92121, USA (email: mitch@liai.org). ²Howard Hughes Medical Institute and Department of Immunology, 1959 NE Pacific Street, University of Washington, Seattle, Washington 98102, USA (email: aruden@u.washington.edu).

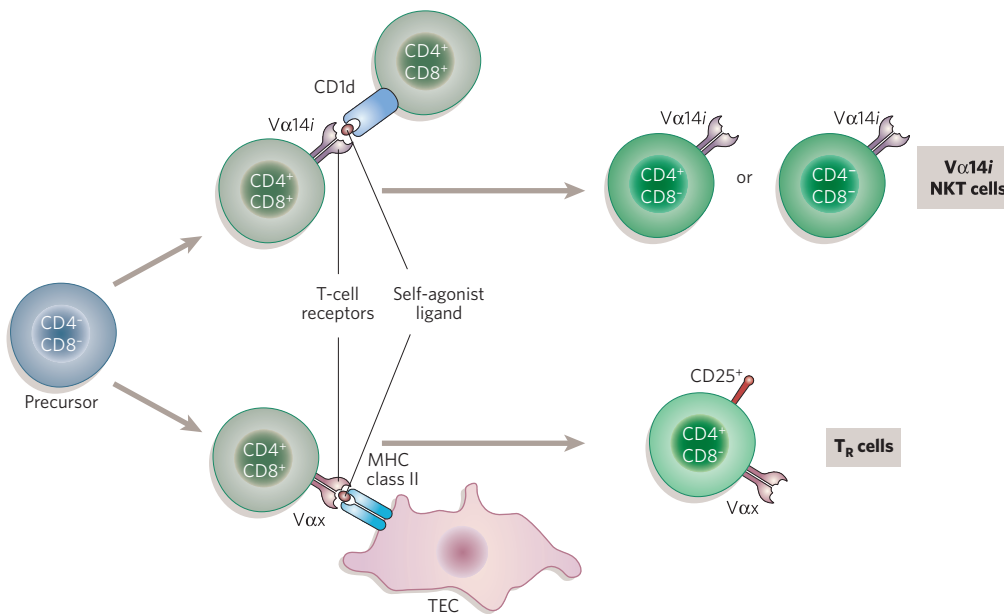


Figure 1 | Recognition of self-agonist ligands in the thymus can create at least two different sublineages of self-reactive T cells. They probably branch off from the mainstream pathway of development at the double-positive stage of differentiation. Thymic T_R precursors can also branch off at the CD4 single-positive stage of differentiation. MHC class II⁺ bone-marrow-derived cells may also participate in T_R-cell selection. TEC, thymic epithelial cell. Vαα, diverse Vα regions.

CD25⁺CD4⁺ T cells, which develop naturally in uninfected healthy individuals, are readily detectable in the thymus and secondary lymphoid organs in mice, rats and humans, where they make up 2–10% of the total CD4⁺ T cells.

Importantly, T_R cells are functionally competent when isolated *ex vivo*. Upon TCR crosslinking, thymic and peripheral CD25⁺CD4⁺ T_R cells suppress proliferation and IL-2 production by responder CD25⁻CD4⁺ or CD8⁺ T cells in a contact-dependent manner. T_R cells themselves have a reduced capacity to proliferate and produce IL-2 or pro-inflammatory cytokines under these conditions^{2,6}. However, this apparent anergy (hyporesponsiveness) ascribed to naturally arising T_R cells is probably an *in vitro* artefact. Adoptive cell transfer and labelling of cycling cells revealed that these cells proliferate *in vivo* and survive over extended periods of time, thereby showing a capacity for self renewal^{7–9}. The robust proliferation of CD25⁺CD4⁺ T cells expressing a transgenic TCR upon stimulation with the cognate ligand is in agreement with these results^{10–12}. Importantly, upon expansion, T_R cells maintain and even enhance their suppressive capacity after proliferation. These results, in combination with the thymic origin of T_R cells, suggest that these cells are a dedicated lineage. Although other T-cell populations may play an analogous suppressive role, the genetic mechanisms underlying their generation and function have not been ascertained, and their dedicated suppressive function has not been proved. We focus our attention on the so-called CD25⁺CD4⁺ T_R cells because of their proven *in vivo* role in suppressing autoimmunity, tumour immunity, allergy and immunity to chronic infections.

Control of immune responses by T_R cells

Several mechanisms might explain T_R-cell-mediated suppression *in vivo*. In particular, expression of the high-affinity trimeric IL-2 receptor by T_R cells might result in growth-factor competition and soak up IL-2 (refs 13, 14). This is unlikely to be a major mechanism because, even in the presence of exogenous IL-2, T_R cells inhibit the upregulation of IL-2 messenger RNA (mRNA) in responder T cells¹⁵. Unlike passive IL-2 deprivation, two major immunosuppressive cytokines (IL-10 and TGF-β) have been implicated as an active effector mechanism of T_R-cell-mediated suppression *in vivo*^{2,16,17}. However, other cell types, including non-regulatory T cells, produce these cytokines. Thus, the relative contribution of TGF-β and IL-10 derived from T_R cells or from other cellular sources to the overall control of autoimmune inflammation in a particular organ and genetic background is currently not known.

In vitro studies show that T_R cells can suppress CD4⁺ and CD8⁺ T-cell responses by an undefined contact-dependent, but IL-10/TGF-β-independent, mechanism⁶. One such mechanism may involve 'reverse' signalling by B7 molecules upon crosslinking on the surface of dendritic cells or activated T cells by CTLA4, the high-affinity receptor for B7 that is expressed at a high level by T_R cells^{18,19}. In dendritic cells, B7 crosslinking induces indoleamine-2,3-dioxygenase, resulting in local tryptophan depletion. In T cells, the biochemical consequences of B7 engagement by CTLA4 remain unclear^{19,20}. In addition, recent reports suggest that pre-activated murine T_R cells induce granzyme-B-dependent, yet perforin-independent, apoptosis in responder T cells²¹. In humans, similarly activated T_R cells are capable of killing responder T cells in a perforin- and granzyme-dependent fashion²². These and other putative contact-dependent mechanisms have been illuminated by *in vitro* studies using antibody-mediated crosslinking of T_R TCR either before or during the suppression assay. The latter may not accurately reflect the physiological level of activation of T_R effector function. Furthermore, the significance of B7 reverse signalling and granzyme/perforin-induced apoptosis/killing of responder T cells as T_R-cell suppression mechanisms operating *in vivo* is not known. Thus, the interplay between cell-contact-dependent and cytokine-dependent mechanisms, the extent of their redundancy and their relative contribution to prevention of autoimmune disease in a specific tissue or organ need to be explored further. In addition, T_R cells can suppress tumour immunity and immune responses that are associated with chronic exposure to infectious agents or environmental antigens².

Control of T_R-cell lineage specification by Foxp3

Although cell-surface expression of CD25 has been instrumental in the isolation and enumeration of T_R cells, its use as a T_R-cell marker during an ongoing immune response is very limited, as all activated CD4⁺ and CD8⁺ T cells transiently upregulate CD25 expression. The search for a definitive T_R-cell marker in mice resulted in the identification of the transcription factor Foxp3, which is expressed in T_R cells but not in recently activated or resting T cells^{23–25}. Several years ago, mutations in the X-chromosome-encoded *Foxp3* gene were identified as the cause of the early-onset fatal autoimmune disorder observed in human IPEX patients (immune dysregulation, polyendocrinopathy, enteropathy, X-linked syndrome) and in *scurfy* mutant mice that spontaneously develop autoimmune disease^{26–28}. In both humans and mice, the various manifestations of autoimmunity are observed in mutant

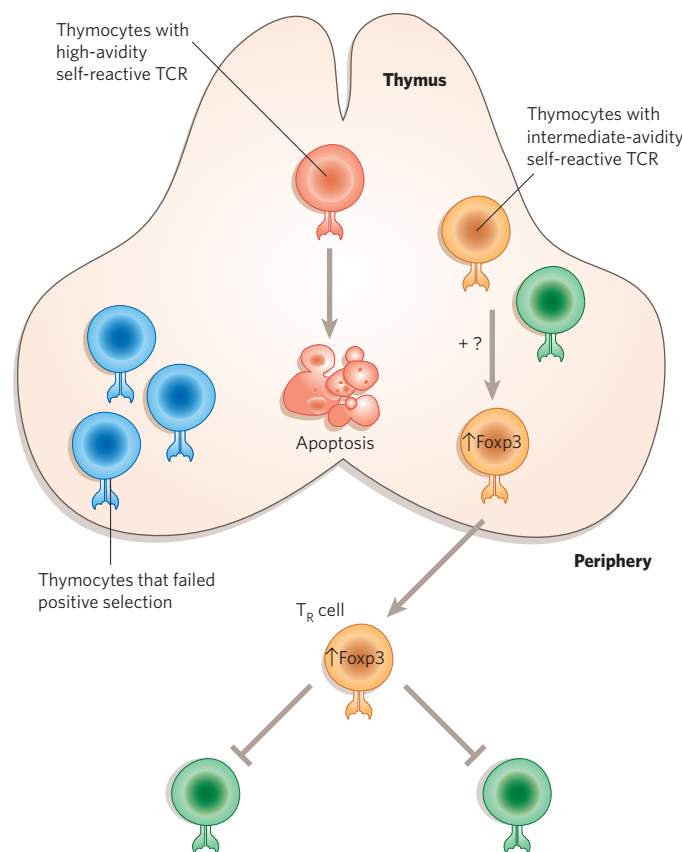


Figure 2 | Alternative fates for developing thymocytes. During development, thymocytes with a high affinity for self-peptide–MHC complexes are deleted (red cells), whereas thymocytes with TCRs that do not react to self undergo death by neglect (blue cells). Thymocytes with a low avidity for self-peptide–MHC are positively selected into the conventional $CD4^+$ and $CD8^+$ lineages. Some self-reactive thymocytes with an intermediate avidity for self-peptide ligands upregulate Foxp3 in response to increased strength or duration of a TCR signal in combination with an unknown signal (yellow cells). Upon Foxp3 induction, thymocytes commit to the T_R -cell lineage and are therefore capable of keeping other T-cell responses in check, thereby preventing autoimmunity.

males, but not in heterozygote female carriers, although the X chromosome undergoes random inactivation in T cells. Therefore approximately 50% of T cells in females lack Foxp3 expression. The resistance of females to autoimmunity is consistent with the ability of T_R -cell-mediated control of immune tolerance to act *in trans* on cells that lack Foxp3. Furthermore, examination of chimaeric mice containing a mixture of Foxp3-deficient and wild-type haematopoietic precursor cells unequivocally demonstrated that $CD25^+CD4^+$ T cells fail to develop from Foxp3-deficient progenitors. Wild-type precursors, however, gave rise to a normal T_R -cell population that could prevent development of overt autoimmune symptoms²³. As a corollary to these results, an expanded $CD25^+CD4^+$ T_R subset is found in mice with T-cell-specific expression of a Foxp3 transgene. Even the $CD25^-CD4^+$ and $CD8^+$ T cells in these mice exhibit some suppressive capacity, as assessed by an *in vitro* assay²⁵. Similarly, retroviral Foxp3 gene transfer into peripheral mouse and human $CD25^-CD4^+$ T cells results in acquisition of suppressive function by at least some of the transduced T cells^{23,24,29,30}. Together, these studies reveal a principal role for Foxp3 in guiding regulatory $CD25^+CD4^+$ T-cell development and function (Fig. 2).

Recent analysis of mice expressing green fluorescent protein ‘knocked into’ the Foxp3 locus suggests that Foxp3 is a T_R -cell-lineage specification factor. This idea stems from the observation that Foxp3 expression is uniquely restricted to a subset of peripheral and thymic

$\alpha\beta$ T cells primarily composed of $CD25^+$ and $CD25^-CD4^+$ T cells with potent suppressive activity. These cells also share a transcriptional signature that is distinct from either $CD25^+Foxp3^-$ or $CD25^-CD4^+Foxp3^-$ T cells³¹. This suggests that Foxp3 is a dedicated genetic mechanism for the generation of T cells that can promote dominant tolerance.

Autoreactivity and T_R -cell lineage commitment

The idea of T_R specificity for self antigens, one of the major tenets in our current understanding of the biology of T_R cells, can be traced to early studies of protection against day-3 thymectomy-induced autoimmunity in specific target tissues, such as the ovaries or thyroid gland^{2,32,33}. Important clues suggesting the requirement for a specific TCR signal for thymocyte commitment to the T_R -cell lineage came from studies of mice expressing a TCR specific for a peptide derived from myelin basic protein (MBP). MBP is a major antigen in the pathogenesis of experimental allergic encephalomyelitis (EAE). This MBP peptide ligand has a low affinity for the MHC class II molecule that presents it, and it is not expressed in the thymus at a level sufficient to induce negative selection of MBP-reactive TCR transgenic T cells. Spontaneous disease was observed only in TCR transgenic mice on the *Rag*^{-/-} background. (Recombination activating genes 1 and 2 (*Rag1*, *Rag2*) encode proteins that mediate recombination in pre-B cells and thymocytes, leading to the production of antibodies and TCRs, respectively). Protection against EAE in these experiments was provided by a regulatory $CD4^+$ T-cell subset that was dependent on the rearrangement of endogenous TCR genes for its development³⁴. These and other TCR transgenic mice lack $CD25^+CD4^+$ T cells in the absence of RAG function. This suggests that only certain TCR specificities support T_R -cell development. An important role for high-affinity TCR engagement in the thymus for T_R -cell development was suggested by an increase in the proportion of $CD25^+CD4^+$ T cells when a transgenic TCR was paired with its cognate ligand, encoded by another transgene^{35,36}.

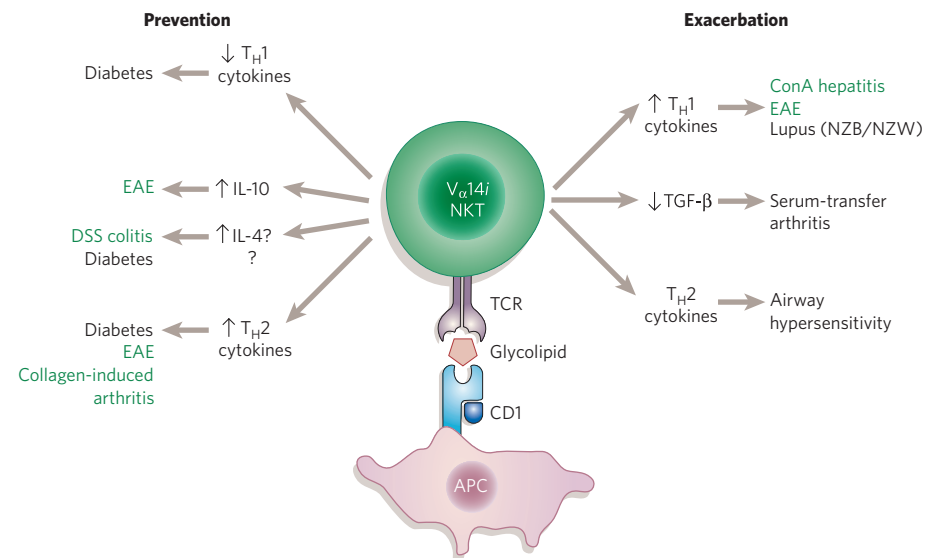
For naturally arising $CD25^+CD4^+$ T_R cells, an increased avidity of TCRs for self-peptide–MHC class II complexes was revealed by the analysis of TCRs from T_R cells compared with non- T_R TCRs upon their transduction into TCR transgenic RAG-deficient T cells of unrelated specificity³⁷. Increased avidity of the TCR for self-peptide–MHC results in a high level of expression of several proteins that attenuates TCR signalling (such as CTLA4, PD-1 and CD5) and increases the level of proteins that promote survival in response to TCR ligation (such as glucocorticoid-induced TNFR-family-related receptor (GITR), TNF-R-II, OX40 and 4BB-1; refs 2, 6, 8). The highly diverse TCR repertoire displayed by naturally arising $CD25^+CD4^+$ T_R cells³⁷ suggests that they recognize a great variety of ligands.

A noticeable skewing of $CD4^+$ T-cell development towards the $CD25^+CD4^+$ T_R -cell lineage observed upon transduction of bone-marrow stem cells with $CD25^+$ T_R -derived, but not $CD25^-$ non- T_R -cell-derived, TCRs suggests that self-reactive TCRs instruct T_R lineage commitment³⁷. An alternative model that suggests that self-reactive TCRs are not responsible for T_R cell lineage commitment has recently challenged this view. Instead, self-reactive TCRs may control the relative size of the T_R subset by affecting expansion or survival of these cells and enabling them to compete with non- T_R cells³⁸. Nevertheless, recent demonstration of an absolute MHC dependence of Foxp3 expression in the Foxp3⁺ subset of immature $CD4^+CD8^+$ double-positive thymocytes, similar to that observed in mature single-positive thymocytes, supports an instructive role for TCR signalling in T_R -cell lineage commitment.

The nature of the antigen-presenting cells (APCs) involved in self-ligand presentation to developing T_R cell precursors in the thymus is not known. However, MHC class II ligand expression by thymic stromal cells is sufficient for the thymic generation of $CD25^+CD4^+$ T_R cells of a particular specificity³⁹.

It is likely that an additional factor(s) cooperates with TCR signals to induce Foxp3 expression in developing thymocytes. Such a factor may be developmentally regulated, as suggested by the inefficient gen-

Figure 3 | Illustration of some of the effects of the activation of V α 14i NKT cells on animal models of autoimmunity and inflammation. Green text indicates that the effect was seen only after addition of exogenous glycolipids such as α GalCer. The black text indicates a role detected in the absence of such stimulation. Autoimmunity can be prevented by T_H2 cytokines by an unknown mechanism, by inhibiting T_H1 cytokine secretion and T_H1-cell expansion in some diabetes experiments or by inducing IL-10 in one EAE (animal model of multiple sclerosis) study. V α 14i NKT cells also exacerbate autoimmunity by stimulating secretion of either T_H1 or T_H2 cytokines or by decreasing TGF- β levels.



eration of CD25⁺CD4⁺ T_R cells in the thymus before day 3 after birth. The absence of these cells is responsible for the aforementioned autoimmunity in day-3 thymectomized mice⁴⁰. CD28 and IL-2R are unlikely to play an indispensable role in thymic Foxp3 induction, despite their major effect on the size of the T_R-cell subset^{41–43}. Therefore, we favour a model of Foxp3 induction and T_R-cell lineage commitment in developing thymocytes in response to an unidentified factor and TCR engagement by self-peptide–MHC complexes within a certain increased range in TCR avidity above that required for conventional T cells (Fig. 2).

In addition to the thymic generation of T_R cells, peripheral non-T_R cells can acquire Foxp3 expression and convert to T_R cells *in vivo* upon chronic antigenic stimulation or under lymphopenic conditions^{44,45}. Thus, acquisition of Foxp3 expression by peripheral non-T_R cells *in vivo* is also facilitated by chronic TCR stimulation and probably by the cytokine environment.

Is autoimmunity associated with Foxp3 mutation due to T_R-cell deficiency?

Lack of T_R cells in Foxp3-deficient mice is the cause of the lymphoproliferative syndrome. Evidence for this is provided by the substantial reduction in this syndrome upon neonatal transfer of a small number of wild-type T_R cells²³. It is noteworthy that the antigen-specific responses of mouse non-regulatory CD4⁺ and CD8⁺ T cells expressing or lacking the Foxp3 gene are indistinguishable when measured by clonal expansion and cytokine production *in vivo* and by cognate peptide-dose response and co-stimulation dependence *in vitro*. Importantly, temporal Foxp3 upregulation is not observed in these cells during the course of conventional antigen-specific immune responses. These observations, combined with the identical onset and progression of autoimmune disease in mice with germline and $\alpha\beta$ T-cell-specific ablation of the Foxp3 gene, provide further proof that T_R-cell deficiency in mice results in a major breakdown of self tolerance and autoimmune pathology affecting multiple organs³¹.

Because the human and murine Foxp3 genes have a high level of sequence conservation throughout the coding and non-coding regions, it is likely that results obtained in mouse studies can be extrapolated to humans. Indeed, human CD25⁺CD4⁺ T_R cells have high levels of Foxp3 expression. As a word of caution, we should mention the conflicting reports as to whether human non-regulatory CD25⁺CD4⁺ or CD8⁺ T cells are capable of transient or prolonged Foxp3 upregulation upon TCR stimulation^{30,46}. One difficulty with the interpretation of this type of study is our current inability to analyse Foxp3 expression at a single-cell level. Expansion of Foxp3-expressing CD25⁺ T cells or acquisition of Foxp3 expression by small numbers of T cells 'pre-

committed' to the T_R-cell lineage may account for these observations. Development of a method to detect Foxp3 expression at the single-cell level would help to address this issue and would permit the monitoring of the dynamics of T_R cells in clinical settings.

The lack of T_R cells in IPEX patients, a result of a Foxp3 deficiency, probably leads to autoimmune pathology in a variety of organs. So, manipulation of the number of T_R cells and their suppressive activity could be tailored to develop novel therapeutic approaches for treatment of the more common forms of autoimmunity that are under polygenic control (see below). Experiments in a variety of models of experimental autoimmunity in mice support this possibility. In this regard, a better understanding of the signals that induce Foxp3, identification of Foxp3 downstream targets and further elaboration of T_R-effector mechanisms are of immediate importance.

Natural killer T cells are a distinct T-cell sublineage

A subset of T cells express receptors found on natural killer cells and are known as NKT cells³. Like T_R cells, NKT cells are a self-reactive T-cell sublineage generated in the thymus (Fig. 1), and they may regulate autoimmunity. However, their developmental pathway is distinct from T_R cells. The specificity of NKT cells is focused on a few antigens and, unlike the CD4⁺CD25⁺ T_R cells, their roles in autoimmunity include both protection from pathogenesis and enhancement of disease⁴⁷.

In mice, the majority of NKT cells are CD4⁺ or double-negative (CD4[−]CD8[−]) T cells that recognize glycolipids presented by CD1d³, a non-classical MHC class I-like antigen-presenting molecule. Most of these cells express a V α 14/J α 18 TCR rearrangement with an invariant CDR3 region. Therefore, they are sometimes referred to as V α 14 invariant (V α 14i) NKT cells or *i*NKT cells⁴⁸ to distinguish them from other T cells that express NK receptors. Most studies indicating a role for NKT cells in autoimmunity have implicated V α 14i NKT cells.

Specificity of *i*NKT cells

The ligand most widely used for activating V α 14i NKT cells is the glycolipid α -galactosylceramide (α GalCer), which was originally isolated from a marine sponge in a screen for compounds that could prevent tumour metastasis⁴⁹. α GalCer binds to CD1d, and the resulting complex is a very strong agonist that binds avidly to the V α 14i TCR^{50–52}. Humans and monkeys also have T cells that are reactive for α GalCer presented by CD1d, and these cells express the TCR V regions orthologous to mouse V α 14 and V β 8, including an invariant V α 24/J α 18 (V α 24i) rearrangement. *i*NKT cells are cross-reactive for APCs expressing either mouse or human CD1d. The strict conservation of this specificity is indicative of its fundamental importance.

Glycosphingolipids closely related to α GalCer⁵³ are abundant in *Sphingomonas* bacteria, which are Gram-negative organisms that lack lipopolysaccharide (LPS). These bacteria are ubiquitous in the environment, and their glycolipids provide a unique example of a microbial glycolipid that can activate the majority of mouse and human *i*NKT cells. Moreover, mice lacking V α 14*i* NKT cells have a reduced ability to clear the bacteria^{54,55}. This information led to the speculation that the rapid immune response of V α 14*i* NKT cells to *Sphingomonas* glycolipids might be analogous to the Toll-like receptor 4 (TLR4)-mediated response to LPS in providing an innate-type reaction important for host defense to bacteria that lack LPS.

Selection of V α 14*i* NKT cells

Although V α 14*i* NKT cells are derived from a double-positive precursor⁵⁶, they branch off from the mainstream pathway of thymic differentiation. Unique features of this pathway include positive selection mediated by CD1d⁺ double-positive thymocytes⁵⁷ (Fig. 1), rather than the cortical epithelial cells required for selection of conventional T cells. A distinct set of factors required for V α 14*i* NKT-cell development is also required. These include cytokines such as IL-15, transcription factors such as NF- κ B family members and T-box expressed in T cells (T-bet), and signalling molecules such as the Src family kinase Fyn⁵⁰ and the adaptor signalling lymphocyte activation molecule-associated protein (SAP) that interacts with Fyn^{58–60} (Table 1). Thymus differentiation programmes the unique properties of V α 14*i* NKT cells, including expression of activation markers and the ability to produce cytokines such as IL-4 and interferon (IFN)- γ ⁵⁰. Mature V α 14*i* NKT cells express inhibitory receptors of the Ly49 family, and these receptors are acquired at a late stage of thymus ontogeny or after export from the thymus. Both the differentiation and immune response of V α 14*i* NKT cells might be regulated by the interplay of TCR signals and inhibitory signals from NK receptors⁶¹, with the inhibitory NK receptors perhaps blocking uncontrolled autoreactive responses.

Isoglobotrihexosylceramide (iGb3), a glycosphingolipid with the structure Gal α (1,3)Gal β (1,4)Glu β (1,1)ceramide, is required for the positive selection of V α 14*i* NKT cells⁶². iGb3 can also stimulate mature *i*NKT cells, indicating that a self agonist can positively select V α 14*i* NKT cells. In contrast to T_R cells, which react with a wide range of self antigens, the invariant TCR of *i*NKT cells is selected to recognize a very limited set of antigens presented by CD1d. The focused self reactivity of *i*NKT cells could have a dual purpose. Presentation of iGb3 in the periphery signals for cellular stress or provides a signal that is important for the immune system to recognize. At the same time, the selection for this specificity might provide an important component of host defence from certain types of bacteria, through recognition of unusual glycosphingolipids.

Regulation of *i*NKT-cell cytokine production

Once activated through their TCR with α GalCer, V α 14*i* NKT cells produce a mixture of T_H1 cytokines, such as TNF and IFN- γ , and T_H2 cytokines, including IL-4 and IL-13, within hours⁴⁷. The rapid production of cytokines, and the speed and intensity of the ensuing activation of dendritic cells and NK cells and other cell types are reminiscent of innate responses. Despite producing a mixture of T_H1 and T_H2 cytokines after TCR activation, V α 14*i* NKT cells can in some cases polarize the immune response in either a T_H1 or T_H2 direction⁴⁷. The modulation of V α 14*i* NKT-cell cytokine production is crucial for their regulation of autoimmunity, although the mechanisms that determine the cytokine polarity are not well understood. The immediate response of V α 14*i* NKT cells to TCR stimulation with agonists such as α GalCer is clearly less easily polarized in either a T_H1 or a T_H2 direction than the response of conventional CD4⁺ T cells to peptide antigen stimulation. The sustained cytokine response of V α 14*i* NKT cells may be influenced, however, by the nature of the glycolipid antigen and by the context in which it is presented. For example, presentation of α GalCer pulsed on dendritic cells favours IFN- γ over IL-4

production⁶³, and the use of altered glycolipid ligands related to α GalCer may help to polarize the V α 14*i* NKT-cell response in either a T_H1 or a T_H2 direction⁶⁴. Additionally, cytokine production by V α 14*i* NKT cells may be determined by the integration of signals from different types of receptor. For example, IL-12 can selectively stimulate IFN- γ production by V α 14*i* NKT cells, in conjunction with the recognition of self antigens⁶⁵.

*i*NKT cells and autoimmunity

Several animal models indicate that V α 14*i* NKT cells prevent autoimmunity and inflammation, either when activated naturally or when using α GalCer or related compounds. Some of these results are summarized in Fig. 3. Stimulation of V α 14*i* NKT cells was beneficial in murine models of diabetes, EAE (which is an animal model of multiple sclerosis), and collagen-induced arthritis⁴⁷. Stimulating these cells was also beneficial in a chemically induced model of colitis. Moreover, the number of V α 14*i* NKT cells is reduced in diabetes-prone non-obese diabetic (NOD) mice, and increasing the number of NKT cells by adoptive transfer, or by expression of a V α 14/J α 18 transgene, reduced this disease⁴⁷. In some studies, V α 14*i* NKT-cell-deficient, CD1^{-/-} NOD mice developed accelerated disease, but this was not always the case⁶⁶, illustrating the controversy surrounding some of the findings on V α 14*i* NKT cells and autoimmunity. In EAE and diabetes, the beneficial effect of stimulating V α 14*i* NKT cells could be attributed in some studies to the induction of IL-4 synthesis by autoantigen-reactive T cells⁴⁷. This was not always observed⁶⁷, and in one study activation of V α 14*i* NKT cells could induce IL-10 synthesis by the autoreactive T cells⁶⁸. Moreover, some protocols for stimulating V α 14*i* NKT cells with synthetic glycolipids exacerbated EAE rather than diminished it⁶⁹. An additional example of V α 14*i* NKT cells promoting autoimmunity is provided by the spontaneous model of systemic lupus erythematosus that arises in (NZB $\times$ NZW)F₁ mice⁷⁰. V α 14*i* NKT cells apparently stimulated pathogenic anti-DNA antibody production, even without activation by exogenous glycolipids. Furthermore, in airway hypersensitivity experiments that model asthma, two reports indicate that mice lacking V α 14*i* NKT cells are resistant to developing hypersensitivity, and IL-4 or IL-13 production by these cells was required for susceptibility^{71,72}.

In parallel with some of the mouse studies, there is evidence that a decrease in V α 14*i* NKT cells in human peripheral blood is correlated with a variety of organ-specific and systemic autoimmune diseases⁷³. Also, when *in vitro* expanded V α 14*i* NKT cells were studied, increased T_H2 cytokine release by these cells correlated with disease that was in remission⁷⁴. In diabetics, however, these conclusions are controversial⁷⁵, and the intrinsic problems of these studies include the variability in the number of V α 14*i* NKT cells in normal humans and assessment of the peripheral blood rather than the site of disease.

In summary, although there are exceptions in which V α 14*i* NKT cells induces IL-10 synthesis⁷⁶ or anergy, in most studies they activate rather than suppress the immune response. This is exemplified by recent results in the K/BxN transgenic mouse model of arthritis in which transfer of serum from these donors causes an arthritis in which deposition of glucose-6-phosphate-isomerase-specific autoantibodies leads to an inflammatory cascade that includes mast cells and complement activation⁷⁷. Recent work implicates V α 14*i* NKT cells in arthritis pathogenesis in this animal model: the level of IFN- γ and IL-4 mRNA decreased in the joints of CD1d^{-/-} mice given the arthritogenic serum, and the level of mRNA for the immunosuppressive cytokine TGF- β 1 increased⁷⁸. This is the opposite of the effect that would be expected for the T_R cells, and immune activation may reflect the true physiological role mediated by V α 14*i* NKT cells in response to infection and in other contexts. Therefore, in many of the cases in which the activity of *i*NKT cells was beneficial in autoimmunity, it might have been through the induction of immune deviation or a cytokine response highly skewed in either a T_H1 or a T_H2 direction, rather than through the induction of a truly unresponsive state.

Conclusions

It has long been believed that the thymus dictates one of three fates to developing T cells: death by neglect, positive selection or negative selection. Recent evidence hints at a 'fourth way'⁷⁹, in which selection by self agonists leads to the differentiation of T-cell subsets with an activated phenotype and specialized functions in the maintenance of self tolerance. The two populations discussed here, however, exhibit different phenotypes, localization and roles in immune regulation. Self recognition is common to both cells, but the factors in the thymus that imprint the distinct fates of T_R cells and Vα14i NKT cells remain to be fully defined.

There are some potential advantages to using T_R cells or Vα14i NKT cells in immune therapy for autoimmune disease, including the fact that the difficult identification of specific autoantigens involved in disease pathogenesis would not be required. Human Vα14i NKT cells can be expanded dramatically by stimulation of their TCR in the presence of cytokines^{80,81}, and the same is true for mouse T_R cells⁸². It may be possible to carry out *ex vivo* expansion of these types of T cell followed by re-infusion of the expanded cells back to patients. These strategies are likely to be cumbersome and expensive, and their practicality remains to be demonstrated. Realization of the therapeutic potential of T_R cells and Vα14i NKT cells probably depends upon devising ways to augment their activity *in vivo* or, in some cases for Vα14i NKT cells, to antagonize their activity.

For Vα14i NKT cells, antagonist or activating glycolipids might have clinical applications, as αGalCer is not toxic⁸³. Antagonist ligands have not been identified, however, and glycolipids that skew the immune response mediated by activated mouse Vα14i NKT cells in a T_H1 or a T_H2 direction have not been extensively tested in humans. An even more daunting problem is the unpredictable nature of the contrasting effects NKT cells have on different autoimmune diseases. Therefore, despite some promise, significant obstacles must be overcome before the therapeutic potential of this cell type can be realized.

The latest clinical studies suggest that treatment of type 1 diabetes patients with non-depleting anti-CD3 antibody induces differentiation of some peripheral T cells into T_R cells or expands the pre-existing T_R subset. The latter may account for the beneficial therapeutic effect of this experimental treatment⁸⁴. In addition to these approaches, recent experimental work has suggested that *Foxp3* gene expression in human T cells specific for a particular autoantigen, or induction of *Foxp3* expression upon activation of autoantigen-reactive, conventional CD4⁺ T cells in the presence of TGF-β, may provide an additional way of generating antigen-specific T_R cells^{23,24,29,30,85,86}. Finally, understanding the critical role of *Foxp3* in T_R biology suggests that this transcription factor can be used as a target for drug development to modulate its expression *in vivo* and thereby affect numbers and activity of T_R cells. Nevertheless, major work is also required to reduce the wealth of knowledge accumulated in experimental models of T_R-mediated inhibition of autoimmune, tumour- and pathogen-specific immune responses to clinical practice. ■

- Janeway, C. A. Jr Approaching the asymptote? Evolution and revolution in immunology. *Cold Spring Harb. Symp. Quant. Biol.* **54**, 1–13 (1989).
- Sakaguchi, S. Naturally arising CD4⁺ regulatory T cells for immunologic self-tolerance and negative control of immune responses. *Annu. Rev. Immunol.* **22**, 531–562 (2004).
- Bendelac, A., Rivera, M. N., Park, S. H. & Roark, J. H. Mouse CD1-specific NK1 T cells: development, specificity and function. *Annu. Rev. Immunol.* **15**, 535–562 (1997).
- Treiner, E. *et al.* Selection of evolutionarily conserved mucosal-associated invariant T cells by MR1. *Nature* **422**, 164–169 (2003).
- Sakaguchi, S., Sakaguchi, N., Asano, M., Itoh, M. & Toda, M. Immunologic self-tolerance maintained by activated T cells expressing IL-2 receptor α-chains (CD25). Breakdown of a single mechanism of self-tolerance causes various autoimmune diseases. *J. Immunol.* **155**, 1151–1164 (1995).
- Shevach, E. M. CD4⁺ CD25⁺ suppressor T cells: more questions than answers. *Nature Rev. Immunol.* **2**, 389–400 (2002).
- Annacker, O. *et al.* CD25⁺ CD4⁺ T cells regulate the expansion of peripheral CD4 T cells through the production of IL-10. *J. Immunol.* **166**, 3008–3018 (2001).
- Gavin, M. A., Clarke, S. R., Negrou, E., Gallegos, A. & Rudensky, A. Homeostasis and anergy of CD4⁺CD25⁺ suppressor T cells *in vivo*. *Nature Immunol.* **3**, 33–41 (2002).
- Fisson, S. *et al.* Continuous activation of autoreactive CD4⁺ CD25⁺ regulatory T cells in the steady state. *J. Exp. Med.* **198**, 737–746 (2003).
- Klein, L., Khazaie, K. & von Boehmer, H. *In vivo* dynamics of antigen-specific regulatory T cells not predicted from behavior *in vitro*. *Proc. Natl Acad. Sci. USA* **100**, 8886–8891 (2003).
- Tarbell, K. V., Yamazaki, S., Olson, K., Toy, P. & Steinman, R. M. CD25⁺ CD4⁺ T cells, expanded with dendritic cells presenting a single autoantigenic peptide, suppress autoimmune diabetes. *J. Exp. Med.* **199**, 1467–1477 (2004).
- Walker, L. S., Chodos, A., Eggena, M., Doms, H. & Abbas, A. K. Antigen-dependent proliferation of CD4⁺ CD25⁺ regulatory T cells *in vivo*. *J. Exp. Med.* **198**, 249–258 (2003).
- Barthlott, T., Kassiotis, G. & Stockinger, B. T cell regulation as a side effect of homeostasis and competition. *J. Exp. Med.* **197**, 451–460 (2003).
- Barthlott, T. *et al.* CD25⁺CD4⁺ T cells compete with naive CD4⁺ T cells for IL-2 and exploit it for the induction of IL-10 production. *Int. Immunol.* **17**, 279–288 (2005).
- Thornton, A. M., Donovan, E. E., Piccirillo, C. A. & Shevach, E. M. Cutting edge: IL-2 is critically required for the *in vitro* activation of CD4⁺CD25⁺ T cell suppressor function. *J. Immunol.* **172**, 6519–6523 (2004).
- Asseman, C., Mauze, S., Leach, M. W., Coffman, R. L. & Powrie, F. An essential role for interleukin 10 in the function of regulatory T cells that inhibit intestinal inflammation. *J. Exp. Med.* **190**, 995–1004 (1999).
- Powrie, F., Carlino, J., Leach, M. W., Mauze, S. & Coffman, R. L. A critical role for transforming growth factor-β but not interleukin 4 in the suppression of T helper type 1-mediated colitis by CD45RB^{low} CD4⁺ T cells. *J. Exp. Med.* **183**, 2669–2674 (1996).
- Paust, S., Lu, L., McCarty, N. & Cantor, H. Engagement of B7 on effector T cells by regulatory T cells prevents autoimmune disease. *Proc. Natl Acad. Sci. USA* **101**, 10398–10403 (2004).
- Fallarino, F. *et al.* Modulation of tryptophan catabolism by regulatory T cells. *Nature Immunol.* **4**, 1206–1212 (2003).
- Mellor, A. L. & Munn, D. H. IDO expression by dendritic cells: tolerance and tryptophan catabolism. *Nature Rev. Immunol.* **4**, 762–774 (2004).
- Gondek, D. C., Lu, L. F., Quezada, S. A., Sakaguchi, S. & Noelle, R. J. Cutting edge: Contact-mediated suppression by CD4⁺CD25⁺ regulatory cells involves a granzyme B-dependent, perforin-independent mechanism. *J. Immunol.* **174**, 1783–1786 (2005).
- Grossman, W. J. *et al.* Human T regulatory cells can use the perforin pathway to cause autologous target cell death. *Immunity* **21**, 589–601 (2004).
- Fontenot, J. D., Gavin, M. A. & Rudensky, A. Y. Foxp3 programs the development and function of CD4⁺ CD25⁺ regulatory T cells. *Nature Immunol.* **4**, 330–336 (2003).
- Hori, S., Nomura, T. & Sakaguchi, S. Control of regulatory T cell development by the transcription factor Foxp3. *Science* **299**, 1057–1061 (2003).
- Khattari, R., Cox, T., Yasayko, S. A. & Ramsdell, F. An essential role for Scurfin in CD4⁺CD25⁺ T regulatory cells. *Nature Immunol.* **4**, 337–342 (2003).
- Bennett, C. L. *et al.* The immune dysregulation, polyendocrinopathy, enteropathy, X-linked syndrome (IPEX) is caused by mutations of FOXP3. *Nature Genet.* **27**, 20–21 (2001).
- Brunkow, M. E. *et al.* Disruption of a new forkhead/winged-helix protein, scurfy, results in the fatal lymphoproliferative disorder of the scurfy mouse. *Nature Genet.* **27**, 68–73 (2001).
- Chatila, T. A. *et al.* JM2, encoding a fork head-related protein, is mutated in X-linked autoimmunity-allergic dysregulation syndrome. *J. Clin. Invest.* **106**, R75–R81 (2000).
- Oswald-Richter, K. *et al.* HIV infection of naturally occurring and genetically reprogrammed human regulatory T-cells. *PLoS Biol.* **2**, E198 (2004).
- Yagi, H. *et al.* Crucial role of FOXP3 in the development and function of human CD25⁺CD4⁺ regulatory T cells. *Int. Immunol.* **16**, 1643–1656 (2004).
- Fontenot, J. D. *et al.* Regulatory T cell lineage specification by the forkhead transcription factor Foxp3. *Immunity* **22**, 329–341 (2005).
- Taguchi, O. *et al.* Tissue-specific suppressor T cells involved in self-tolerance are activated extrathymically by self-antigens. *Immunology* **82**, 365–369 (1994).
- Seddon, B. & Mason, D. Peripheral autoantigen induces regulatory T cells that prevent autoimmunity. *J. Exp. Med.* **189**, 877–882 (1999).
- Olivares-Villagomez, D., Wang, Y. & Lafaille, J. J. Regulatory CD4⁺ T cells expressing endogenous T cell receptor chains protect myelin basic protein-specific transgenic mice from spontaneous autoimmune encephalomyelitis. *J. Exp. Med.* **188**, 1883–1894 (1998).
- Jordan, M. S. *et al.* Thymic selection of CD4⁺CD25⁺ regulatory T cells induced by an agonist self-peptide. *Nature Immunol.* **2**, 301–306 (2001).
- Apostolou, I., Sarukhan, A., Klein, L. & von Boehmer, H. Origin of regulatory T cells with known specificity for antigen. *Nature Immunol.* **3**, 756–763 (2002).
- Hsieh, C. S. *et al.* Recognition of the peripheral self by naturally arising CD25⁺ CD4⁺ T cell receptors. *Immunity* **21**, 267–277 (2004).
- van Santen, H. M., Benoist, C. & Mathis, D. Number of T reg cells that differentiate does not increase upon encounter of agonist ligand on thymic epithelial cells. *J. Exp. Med.* **200**, 1221–1230 (2004).
- Modigliani, Y. *et al.* Establishment of tissue-specific tolerance is driven by regulatory T cells selected by thymic epithelium. *Eur. J. Immunol.* **26**, 1807–1815 (1996).
- Asano, M., Toda, M., Sakaguchi, N. & Sakaguchi, S. Autoimmune disease as a consequence of developmental abnormality of a T cell subpopulation. *J. Exp. Med.* **184**, 387–396 (1996).
- Salomon, B. *et al.* B7/CD28 costimulation is essential for the homeostasis of the CD4⁺CD25⁺ immunoregulatory T cells that control autoimmune diabetes. *Immunity* **12**, 431–440 (2000).
- Furtado, G. C., Curotto de Lafaille, M. A., Kutchukhidze, N. & Lafaille, J. J. Interleukin 2 signaling is required for CD4⁺ regulatory T cell function. *J. Exp. Med.* **196**, 851–857 (2002).
- Malek, T. R. & Bayer, A. L. Tolerance, not immunity, crucially depends on IL-2. *Nature Rev. Immunol.* **4**, 665–674 (2004).
- Apostolou, I. & von Boehmer, H. *In vivo* instruction of suppressor commitment in naive T cells. *J. Exp. Med.* **199**, 1401–1408 (2004).
- Curotto de Lafaille, M. A., Lino, A. C., Kutchukhidze, N. & Lafaille, J. J. CD25⁺ T cells generate CD25⁺ Foxp3⁺ regulatory T cells by peripheral expansion. *J. Immunol.* **173**, 7259–7268 (2004).
- Walker, M. R. *et al.* Induction of Foxp3 and acquisition of T regulatory activity by stimulated human CD4⁺CD25⁺ T cells. *J. Clin. Invest.* **112**, 1437–1443 (2003).
- Godfrey, D. I. & Kronenberg, M. Going both ways: immune regulation via CD1d-dependent NKT cells. *J. Clin. Invest.* **114**, 1379–1388 (2004).

48. Godfrey, D. I., MacDonald, H. R., Kronenberg, M., Smyth, M. J. & Van Kaer, L. NKT cells: what's in a name? *Nature Rev. Immunol.* **4**, 231–237 (2004).
49. Morita, M. *et al.* Structure-activity relationship of α -galactosylceramides against B16-bearing mice. *J. Med. Chem.* **38**, 2176–2187 (1995).
50. Kronenberg, M. Towards an understanding of NKT cell biology: Progress and paradoxes. *Annu. Rev. Immunol.* **23**, 877–900 (2005).
51. Benlagha, K., Weiss, A., Beavis, A., Teyton, L. & Bendelac, A. *In vivo* identification of glycolipid antigen-specific T cells using fluorescent CD1d tetramers. *J. Exp. Med.* **191**, 1895–1903 (2000).
52. Matsuda, J. L. *et al.* Tracking the response of natural killer T cells to a glycolipid antigen using CD1d tetramers. *J. Exp. Med.* **192**, 741–754 (2000).
53. Kawahara, K., Kuraishi, H. & Zahring, U. Chemical structure and function of glycosphingolipids of *Sphingomonas* spp and their distribution among members of the $\alpha 4$ subclass of Proteobacteria. *J. Ind. Microbiol. Biotechnol.* **23**, 408–413 (1999).
54. Kinjo, Y. *et al.* Recognition of bacterial glycolipids by natural killer T cells. *Nature* **434**, 520–525 (2005).
55. Mattner, J. *et al.* Exogenous and endogenous glycolipid antigens activate NKT cells during microbial infections. **434**, 525–529 (2005).
56. Gapin, L., Matsuda, J. L., Surh, C. D. & Kronenberg, M. NKT cells derive from double-positive thymocytes that are positively selected by CD1d. *Nature Immunol.* **2**, 971–978 (2001).
57. Coles, M. C. & Raulet, D. H. NK1.1⁺ T cells in the liver arise in the thymus and are selected by interactions with class I molecules on CD4⁺CD8⁺ cells. *J. Immunol.* **164**, 2412–2418 (2000).
58. Nichols, K. E. *et al.* Regulation of NKT cell development by SAP, the protein defective in XLP. *Nature Med.* **11**, 340–345 (2005).
59. Chung, B., Aoukaty, A., Dutz, J., Terhorst, C. & Tan, R. Signaling lymphocytic activation molecule-associated protein controls NKT cell functions. *J. Immunol.* **174**, 3153–3157 (2005).
60. Pasquier, B. *et al.* Defective NKT cell development in mice and humans lacking the adapter SAP, the X-linked lymphoproliferative syndrome gene product. *J. Exp. Med.* **201**, 695–701 (2005).
61. Voyle, R. B. *et al.* Ligand-dependent inhibition of CD1d-restricted NKT cell development in mice transgenic for the activating receptor Ly49D. *J. Exp. Med.* **197**, 919–925 (2003).
62. Zhou, D. *et al.* Lysosomal glycosphingolipid recognition by NKT cells. *Science* **306**, 1786–1789 (2004).
63. Fujii, S., Shimizu, K., Kronenberg, M. & Steinman, R. M. Prolonged IFN γ -producing NKT response induced with α -galactosylceramide-loaded DCs. *Nature Immunol.* **3**, 867–874 (2002).
64. Miyamoto, K., Miyake, S. & Yamamura, T. A synthetic glycolipid prevents autoimmune encephalomyelitis by inducing T $_{H2}$ bias of natural killer T cells. *Nature* **413**, 531–534 (2001).
65. Brigl, M., Bry, L., Kent, S. C., Gumperz, J. E. & Brenner, M. B. Mechanism of CD1d-restricted natural killer T cell activation during microbial infection. *Nature Immunol.* **4**, 1230–1237 (2003).
66. Hammond, K. J. & Kronenberg, M. Natural killer T cells: natural or unnatural regulators of autoimmunity? *Curr. Opin. Immunol.* **15**, 683–689 (2003).
67. Beaudoin, L., Laloux, V., Novak, J., Lucas, B. & Lehuen, A. NKT cells inhibit the onset of diabetes by impairing the development of pathogenic T cells specific for pancreatic beta cells. *Immunity* **17**, 725–736 (2002).
68. Furlan, R. *et al.* Activation of invariant NKT cells by α GalCer administration protects mice from MOG35–55-induced EAE: critical roles for administration route and IFN γ . *Eur. J. Immunol.* **33**, 1830–1838 (2003).
69. Jahng, A. W. *et al.* Activation of natural killer T cells potentiates or prevents experimental autoimmune encephalomyelitis. *J. Exp. Med.* **194**, 1789–1799 (2001).
70. Zeng, D., Liu, Y., Sidobre, S., Kronenberg, M. & Strober, S. Activation of natural killer T cells in NZB/W mice induces Th1-type immune responses exacerbating lupus. *J. Clin. Invest.* **112**, 1211–1222 (2003).
71. Akbari, O. *et al.* Essential role of NKT cells producing IL-4 and IL-13 in the development of allergen-induced airway hyperreactivity. *Nature Med.* **9**, 582–588 (2003).
72. Lisbonne, M. *et al.* Cutting edge: invariant V α 14 NKT cells are required for allergen-induced airway inflammation and hyperreactivity in an experimental asthma model. *J. Immunol.* **171**, 1637–1641 (2003).
73. van der Vliet, H. J. *et al.* Circulating V(α 24⁺) V β 11⁺ NKT cell numbers are decreased in a wide variety of diseases that are characterized by autoreactive tissue damage. *Clin. Immunol.* **100**, 144–148 (2001).
74. Araki, M. *et al.* Th2 bias of CD4⁺ NKT cells derived from multiple sclerosis in remission. *Int. Immunol.* **15**, 279–288 (2003).
75. Lee, P. T. *et al.* Testing the NKT cell hypothesis of human IDDM pathogenesis. *J. Clin. Invest.* **110**, 793–800 (2002).
76. Sonoda, K. H. *et al.* NK T cell-derived IL-10 is essential for the differentiation of antigen-specific T regulatory cells in systemic tolerance. *J. Immunol.* **166**, 42–50 (2001).
77. Matsumoto, I. *et al.* How antibodies to a ubiquitous cytoplasmic enzyme may provoke joint-specific autoimmune disease. *Nature Immunol.* **3**, 360–365 (2002).
78. Kim, H. Y. *et al.* NKT cells promote antibody-induced joint inflammation by suppressing transforming growth factor β 1 production. *J. Exp. Med.* **201**, 41–47 (2005).
79. Baldwin, T. A., Hogquist, K. A. & Jameson, S. C. The fourth way? Harnessing aggressive tendencies in the thymus. *J. Immunol.* **173**, 6515–6520 (2004).
80. Brossay, L. *et al.* CD1d-mediated recognition of an α -galactosylceramide by natural killer T cells is highly conserved through mammalian evolution. *J. Exp. Med.* **188**, 1521–1528 (1998).
81. Rogers, P. R. *et al.* Expansion of human V α 24⁺ NKT cells by repeated stimulation with KRN7000. *J. Immunol. Methods* **285**, 197–214 (2004).
82. Tang, Q. *et al.* *In vitro*-expanded antigen-specific regulatory T cells suppress autoimmune diabetes. *J. Exp. Med.* **199**, 1455–1465 (2004).
83. Giaccone, G. *et al.* A phase I study of the natural killer T-cell ligand α -galactosylceramide (KRN7000) in patients with solid tumors. *Clin. Cancer Res.* **8**, 3702–3709 (2002).
84. Bluestone, J. A. & Tang, Q. Therapeutic vaccination using CD4⁺CD25⁺ antigen-specific regulatory T cells. *Proc. Natl Acad. Sci. USA* **101** Suppl 2, 14622–14626 (2004).
85. Rao, P. E., Petrone, A. L. & Ponath, P. D. Differentiation and expansion of T cells with regulatory function from human peripheral lymphocytes by stimulation in the presence of TGF- β . *J. Immunol.* **174**, 1446–1455 (2005).
86. Chen, W. *et al.* Conversion of peripheral CD4⁺CD25⁺ naive T cells to CD4⁺CD25⁺ regulatory T cells by TGF- β induction of transcription factor Foxp3. *J. Exp. Med.* **198**, 1875–1886 (2003).

Acknowledgements We thank members of our laboratories for highly stimulating discussions and support. This work was supported by NIH grants. A.R. is an Investigator with the Howard Hughes Medical Institute.

Competing interests statement The authors declare that they have no competing financial interests.

An array of possibilities for the study of autoimmunity

C. Garrison Fathman¹, Luis Soares^{1,2}, Steven M. Chan¹ & Paul J. Utz¹

Since the completion of the sequencing of the human genome, scientific focus has shifted from studying genes to analysing the much larger number of proteins encoded by them. Several proteins can be generated from a single gene depending on how the genetic information is read (transcribed) and how the resultant protein is modified following translation (post-translational modification). Genomic and proteomic technologies are already providing useful information about autoimmune disease, and they are likely to lead to important discoveries within the next decade.

Autoimmune diseases result from three interacting components: genetic, environmental and regulatory¹ (Fig. 1). Autoimmunity is caused by a complex interaction of multiple gene products, unlike immunodeficiency diseases, where a single dominant genetic trait is often the main disease determinant². High-throughput analysis can tell us which genes are turned on or off in different tissues from patients with autoimmune disease or in cells following different stimuli, but analysis of messenger RNA (mRNA) expression alone is insufficient to determine whether the proteins encoded are synthesized.

Through recently developed technology we can analyse protein expression and use the information gained to complement the gene-expression findings.

These two major investigational approaches, genomic and proteomic, have great potential to shed light on the molecular basis of autoimmune disease in a genetically diverse population. New techniques are likely to markedly accelerate the rate of discovery and characterization of disease-specific genetic and metabolic pathways, and will lead eventually to the development of individualized therapies that take into account markers of disease predisposition and therapeutic response. This review focuses on the main technologies that are being applied to dissect the genome and proteome in autoimmunity. Emphasis is placed on emerging techniques, and the controversial aspects of genomics and proteomics research are discussed.

Genomics and autoimmunity

Traditional methods of differential cloning have been employed successfully to isolate unique genes associated with disease. But these techniques have limited use in the study of multigenic diseases such as autoimmunity. Complementary DNA (cDNA) microarrays, with their ability to determine the expression pattern of thousands of genes simultaneously and to obtain molecular signatures of the state of activity of a cell, are better suited to such studies. It is generally recognized that expression of several genes is coordinated both spatially and temporally and that this coordination changes during the development and progression of disease. Microarray analysis should provide valuable information on disease pathology, progression, response to treatment and overall cellular microenvironments and should also lead to improved, timely diagnosis and novel therapeutic approaches for autoimmune diseases. The genetic datasets obtained, however, are usually highly complex, and the assignment of biological function to the new genes requires other biological methods, such as proteomic analysis, before genetic products can be placed in functional classes or attributed precise roles in cellular pathways.

With the aid of high-throughput sequencing and associated bioinformatic resources, complete sequences of entire genomes, including the human genome (for example, www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Genome or www.ensembl.org), are now

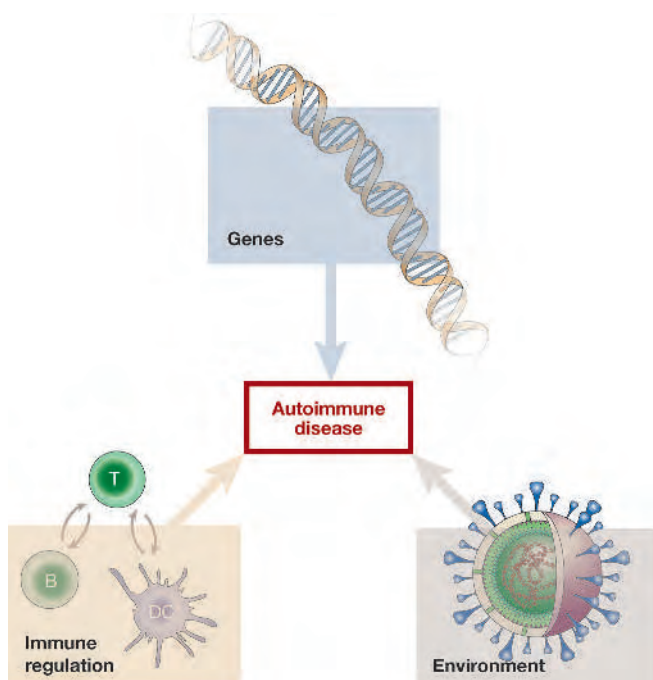


Figure 1 | Requirements for the development of autoimmune disease. The environment can trigger autoimmunity in genetically predisposed individuals under conditions of immune dysregulation.

¹Department of Medicine, Division of Immunology and Rheumatology, Stanford University School of Medicine, Stanford, California 94305, USA (e-mail: cfathman@stanford.edu). ²IBMP, Curitiba, PR81220, Brazil.

available in an organized form within public databases. Concomitant with the development of these new resources, several methods of comparative gene expression have been created to take advantage of the newly available genomic information. Classic quantitative approaches such as northern blotting and RNase protection assays, for which *a priori* sequence information is required, have been essential tools. The recent development of investigational tools such as differential display, serial analysis of gene expression (SAGE)³ and DNA microarrays⁴ will enable the investigation of differences in gene-expression patterns that will increase our understanding of autoimmune disease pathogenesis and the construction of disease-specific 'molecular fingerprinting' models. When coupled to functional proteomics, these methods gain a second necessary biological dimension as the consequences of altered gene expression can be closely evaluated.

cDNA microarray technology

Microarrays come in two basic varieties: arrays where longer DNA fragments are printed onto a solid support, and arrays where short oligonucleotides are synthesized *in situ*. The basic concept of data generation is the same: mRNA is reverse-transcribed into cDNA, labelled with a fluorescent dye and hybridized to the array. After washing away any unbound sample, the array is scanned. The fluorescent intensities at a specific spot representing an individual gene directly correlate with the abundance of this gene in the sample. Several methods and software tools have been developed to handle the large volumes of data generated in microarray experiments⁵. At the simplest level, two samples are compared for differentially expressed genes by calculating the ratios between their fluorescent signals. SAM (significance analysis of microarrays) and other statistical programs allow sophisticated comparison of samples, including assignment of statistical significances to observed differences in gene expression⁶. Another technique is hierarchical clustering, which uses standard mathematical algorithms to cluster genes with a similar expression pattern across all samples (for example, a time-course experiment) into a dendrogram, where increasing distance between branches reflects increasing dissimilarity of gene-expression patterns⁷. This method was used to analyse the response of human fibroblasts (cells thought to play a critical role in autoimmune diseases such as scleroderma⁸) to serum. The most striking finding was the coordinated regulation of expression of genes whose products act at different steps in a common process, such as cell-cycle coordination and proliferation. This suggests that unknown genes within these clusters have a related function. These interrelated gene products, acting together, are called an 'interactome' (see below).

Drawbacks of cDNA microarray technology

Unfortunately, cDNA microarray analysis is in its infancy for the study of autoimmune diseases. Although many reports have described the use of commercial microarrays, all of the current studies are plagued with incomplete and potentially misleading data sets. Despite this, several important observations have been reported. Many of the difficulties stem from the use, in current commercial microarrays, of small oligonucleotides that generally correspond to a single exon of a 'gene'. Thus investigators are generally looking at 'the ends of genes'. Yet the human and mouse genomes are composed of many segments of coding sequence (exons), interspersed with non-coding segments (introns). A recent genome-wide study of human splicing events demonstrated that at least 74% of human multi-exon genes are alternatively spliced⁹.

Another problem is that alternative splicing varies depending on the circumstance. For example, under conditions of inflammatory stress, unique isoforms (splice variants) are the rule¹⁰. Non-canonical alternative splicing may be an important mechanism for the generation of epitopes to which the immune system is not tolerant, which may lead to autoimmune responses. Current microarray platforms for mammalian gene expression do not allow the identification of splice variants; analyses of many datasets may classify a 'gene' as upregulated

when it may in fact be a splice variant, expressing the exon defined by the oligonucleotide, but having actions that may not reflect the 'gene' that was intended. For example, microarray analysis may indicate that the expression of the gene encoding Otubain-1 is upregulated, as all of its potential 13 variants, which have different (and some opposing) functions, share the same exon defined by the microarray probe¹¹. To overcome this potential for misleading data interpretation, several groups have developed exon-specific arrays and arrays composed of whole-nucleotide sequences to obtain true representation of alternative splicing^{9,12}.

Successful adaptation of cDNA microarray analysis to disease

Global patterns of gene expression can be monitored during disease progression and after clinical intervention. cDNA microarray technology has been successfully used in the clinical setting to study disease biology, especially cancer and autoimmune diseases^{13–22}. Golub and colleagues reported that mRNA-transcript profiles for leukaemia cells could be divided into acute myeloid and acute lymphoblastic subtypes without prior knowledge²¹. Alizadeh and colleagues identified distinct signature profiles in diffuse large B-cell lymphomas that correlated with the disease profile²². The Steinman laboratory has used microarray technology for large-scale analysis of mRNA transcripts from complex tissues including human brain specimens from patients with multiple sclerosis¹³. Other studies using gene microarrays and large-scale robotic sequencing of libraries to study brains of patients with multiple sclerosis or Huntington's disease have also been reported^{13–15}. Smaller cDNA arrays have been used in the analysis of gene expression in peripheral blood mononuclear cells of patients with spondyloarthropathy and systemic lupus erythematosus (SLE)^{16,17}. In some cases, important information on the expression of particular clusters of genes, such as anti-apoptotic genes in B-cell lymphomas²², chemokine receptor CXCR4 in arthritis¹⁸ and tumour-necrosis factor (TNF)/death receptor family members and interferon-related gene expression in SLE¹⁷, has been discovered and has opened up new avenues for therapeutic targets.

Cell lines with a pattern of gene expression observed in autoimmune disease can be screened to select potential drug candidates that can restore the 'normal' pattern of expression. Similarly, transgenic animal models expressing or lacking selected genes can be analysed following therapy for modifications to the overall pattern of gene expression that closely resembles that of the affected subject. Gene-expression patterns seen in autoimmune diseases are also reflected in the non-diseased first-degree relatives of the disease probands¹⁹.

Defining the interactome

Following the identification of gene-expression patterns by microarray analysis, precise protein–protein interactions need to be worked out to define the interactome. Scientists working on yeast²³ and *Caenorhabditis elegans*²⁴ have pooled their data to generate functional maps of the interactome. But, at present, many data from mammalian studies of protein–protein interaction are not publicly available, and there has been little attempt to share such data. Until this occurs, we will generally depend on the current 'guilt by association' model of gene interactions in autoimmune diseases²⁵. This model assumes that proteins with correlated expression levels in microarray analyses²⁶ performed under the same series of conditions are functionally linked. True representation of the interactome will require the analysis of functional links between gene clusters using mating-based yeast two-hybrid assays^{23,24}. This system identifies gene products that interact biochemically at the level of the expressed protein product. By choosing protein domains using available information on amino-acid sequence or secondary structure (such as localization signals, transmembrane regions and domain composition), this technique can be used to study intracellular, transmembrane and secreted protein interactions. Briefly, gene (open reading frame) clones from a cluster of co-expressed genes (obtained from collections such as the

Table 1 | Unbiased proteomic technologies with potential applications in the study of autoimmune diseases

Technology	Description	Limitations	Current/potential applications	References
Mass spectrometry	Separates a complex mixture of ionized proteins or peptides on the basis of their mass-to-charge ratio.	Its application in defining autoimmune 'biosignatures' with prognostic or diagnostic significance has not yet been reported. The identities of the peaks generated by the technique are not immediately known.	Identification of serum proteomic patterns that distinguish neoplastic from non-neoplastic disease in the ovary. Identification of biomarkers in the synovial fluid and serum of rheumatoid arthritis patients that correlate with disease severity. Classification of autoimmune patients into prognostic subgroups based on serum 'biosignatures'.	54 31
Two-dimensional gel electrophoresis	Separates a complex mixture of proteins on the basis of molecular mass and isoelectric point.	Not easily reproducible. Not all classes of proteins can be resolved. Limited dynamic range/semi-quantitative.	Detection of aberrant protein expression in bodily fluids, immune cells and/or diseased tissues.	29,35
Whole proteome	Proteins coded by all known genes are expressed, purified and deposited on microarray slides.	Huge investment in resources for the production of purified proteins. Purified proteins may not be correctly folded or appropriately modified (for example, glycosylation and phosphorylation). Not a realistic approach for more complex organisms.	So far, only a yeast whole-proteome microarray has been fabricated. The specificity of purified antibodies was analysed using this array. Rapid identification of novel target autoantigens by probing microarrays with autoimmune serum.	45,55

I.M.A.G.E consortium (<http://image.LLNL.gov>) or by polymerase chain reaction) can be fused in-frame with the DNA-binding domain of GAL4, which represents the bait, or to the activation domain of GAL4, which represents the prey. This is carried out in a haploid yeast strain. The bait and prey pools are systematically mated and the transformants selected for the activation of reporter genes. Positive interactions are catalogued, and the resulting binary data can be used in the construction of an interactome. The interactome is then screened against the public literature, where a function may have been noted for one or more members of the network.

As mentioned above, this methodology has been used successfully in yeast and *C. elegans*^{23,27,28}, as well as in bacteria. Although these organisms have considerably smaller genomes, it should be possible to use this approach to identify functional interactomes in mammalian tissues in autoimmune disease. Interaction mapping of the whole human or mouse genome is not needed; instead, a subset of disease-related genes in relevant tissues can be analysed. For example, inhibition of TNF is beneficial in many autoimmune diseases (see review by Feldmann and Steinman in this issue, page 612), but what is the mechanism for this TNF inhibition? What is the TNF interactome? Does TNF play a direct role in the final common pathway of multiple autoimmune disease effects, or does it act further upstream, disrupting several different pro-inflammatory pathways, each relevant in different therapeutic uses of TNF inhibition? It is only through identification of gene-expression patterns linked to disease pathophysiology, followed by validation of candidate genes through proteomic approaches, that we will realize the success of these new technologies.

Unbiased proteomic technologies

Two major proteomes are the focus of most efforts to understand autoimmune diseases: the serum proteome and the cellular proteome. There are two approaches to investigating each proteome, unbiased and biased. Unbiased technologies attempt to separate and quantify every protein of the expressed proteome (Table 1). This is based on the hypothesis that proteins that are differentially expressed or modified in samples from patients but not in those from healthy controls are likely to be involved in the autoimmune process or at least serve as biomarkers that correlate with disease or therapeutic outcome. Thus, in theory, unbiased approaches have the potential to identify any protein that is involved in autoimmunity without the need for prior knowledge of the protein or its function. In reality, not all proteins are amenable to this kind of analysis because only a fraction of the proteome can be analysed accurately and reliably.

The two most commonly used unbiased technologies are two-dimensional (2D) gel electrophoresis and mass spectrometry (Table 1; ref. 20). With both of these methods, comparisons are made of samples from patients and healthy controls. The main limitations of 2D gel electrophoresis are its low throughput and poor sensitivity, with some estimates suggesting that less than 50% of expressed proteins in cells are amenable to such analysis²⁹. The use of mass spectrometry is plagued by similar, and other, limitations that are beyond the scope of this discussion. An excellent review on this subject has recently been published³⁰.

Despite these limitations, unbiased methods show promise as tools for protein discovery in autoimmunity and have recently led to the identification of candidate markers in the synovial fluid of patients with rheumatoid arthritis³¹ and the cerebrospinal fluid of patients with multiple sclerosis³². In the case of 2D gel electrophoresis, S100A9, a small calcium-binding protein, was identified as a candidate diagnostic marker for rheumatoid arthritis by comparing the protein profiles of synovial fluid from rheumatoid arthritis and osteoarthritis patients³³. In an analogous study using mass spectrometry, C-reactive protein and six members of the S100 protein family (including S100A9) were elevated in the synovial fluid of patients with erosive rheumatoid arthritis compared with that of patients with non-erosive rheumatoid arthritis³¹. In addition, Stone and colleagues analysed, by mass spectrometry, serum from patients with Wegener's granulomatosis who were enrolled in a trial of the drug etanercept (a TNF inhibitor). They showed that this technique could distinguish between patients in stable clinical remission and those with the active disease³⁴. It is important to note that the proteins or peptides identified using unbiased technologies are simply correlated with disease activity, and that their role in disease causation can only be determined through loss- or gain-of-function experiments. NHLBI's Proteomics Consortium has made a major commitment to these two unbiased approaches, with all but one of the ten centres using one or both of these techniques to understand blood diseases (<http://www.nhlbi-proteomics.org/>).

Biased proteomic technologies

Once a serum or cellular component of interest is identified using one of several approaches (for example, the unbiased proteomics assays described above, genomic profiling, genetic studies or analysis of knockout animals), methods that enable more focused studies of these 'known' components are required — which we term biased approaches (Table 2). The problem with biased approaches is that only a limited subset of proteins to which detection reagents, such as

Table 2 | Biased proteomic technologies with potential applications in the study of autoimmune diseases

Technology	Description	Limitations	Current/potential applications	References
Autoantigen microarrays	Purified proteins and peptides of known auto-antigens are deposited on microarray slides and probed with auto-immune serum or other biological fluids.	Only previously defined autoantigens for a particular disease are represented on the array. Significance of autoantibodies in autoimmune progression is controversial for many autoimmune diseases.	Monitoring disease progression and response to DNA vaccine treatment in a murine model of multiple sclerosis by autoantibody profiling. Microarrays produced for the study of autoantibodies in human connective-tissue diseases.	38,40,43,55
Antibody (forward-phase) microarrays	Highly specific antibodies are immobilized on microarray slides. Arrays are probed with cell culture supernatant, cell lysate or serum samples.	Limited usefulness in the detection of intracellular proteins and their phosphorylation events owing to the lack of commercially available antibodies that function in this format. Cross-reactivity of antibodies.	A 51-feature antibody microarray was used to measure cytokine secretion from stimulated dendritic cells. High-throughput identification of cytokines in autoimmune diseases.	56
Bead-based multiplexed	Bead-based assays are similar to autoantigen and antibody microarrays in concept. Instead of a slide surface, proteins are coated on differentially identifiable beads.	Degree of multiplexing limited by the number of differentially identifiable beads. Similar limitations to autoantigen and antibody microarrays.	Bead-based assays are currently dominated by cytokine-detection assays. Autoantigen-coated beads for autoantibody profiling.	47
MHC-tetramer microarrays	MHC-tetramers loaded with different peptides are deposited on a slide surface and probed with T-cell populations.	It is not known whether this technique is sufficiently sensitive to detect autoreactive T cells in autoimmune patients.	Detection of antigen-specific T cells in vaccinated mice. Monitoring the diversity of peptide epitopes targeted by T cells in autoimmune diseases.	46
Reverse-phase protein (lysate) microarrays	A large number of lysate samples are deposited as microspots on a slide surface and probed with phospho-specific or pan-specific antibodies.	A limited number of analytes can be analysed on a single slide even with the use of multi-sectored slides. Cross-reactivity of antibodies.	Differential STAT protein phosphorylation in interleukin-2 stimulated regulatory T cells. Successful demonstration of this technology in defining signalling profiles in laser-capture microdissected cancer samples. Identification of signalling defects in auto-immune cells.	50,58
Multiparameter phospho-protein flow cytometry	Detection of multiple phospho-states in single cells using phospho-specific antibodies and multiparameter flow cytometry.	Not all phospho-epitopes can be detected using this technology. Cross-reactivity of antibodies. Degree of multiplexing limited by the number of differentially identifiable fluorophores.	Classification of patients with acute myeloid leukaemia into clinically relevant subgroups on the basis of their phosphorylation states in response to external stimuli. Identification of signalling defects in auto-immune cells.	51

monoclonal antibodies, have been developed can be analysed. Indeed, one of the biggest challenges facing the field of proteomics is the development of high-quality reagents for detecting a large number of proteins.

The main components in the serum proteome that are of particular interest in autoimmunity are autoantibodies and inflammatory mediators, such as cytokines and chemokines. In the cellular proteome, components of interest include T-cell receptors (TCRs), other cell-surface proteins and intracellular signalling proteins. Although this classification scheme is not comprehensive, it provides a useful framework for understanding how different biased technologies can contribute to our understanding of autoimmunity.

The main biased techniques used for large-scale analysis of many proteins are multiplexed western blots³⁵ and protein microarrays^{36,37}. Western blotting is a time-tested technique used in virtually all disciplines of biology. Straightforward protocols have been developed for studying nearly 800 different analytes simultaneously, taking advantage of the fact that proteins migrate at different molecular weights by gel electrophoresis³³. Several examples of this technique have now been described for studying autoimmune disease, particularly rheumatoid arthritis³³. Lorenz and colleagues compared the proteome in the synovial tissue of patients with rheumatoid arthritis with that of patients with osteoarthritis³³. Comparison of the transcriptome and proteome revealed that only 28% of the mRNA and proteins correlated between the patient groups. Moreover, significant differences at the protein level were noted for Stat1, cathepsin D and p47phox, which may be useful targets for therapy.

The greatest recent advance in proteomics studies in autoimmunity is the use of protein microarrays. Arrays have been developed specifi-

cally for the study of numerous components of both the serum and cellular proteomes, and are described below (Fig. 2).

Antibody profiling using arrays of antigens

Antigen microarrays have been created for a variety of diseases, including infectious diseases (such as HIV; ref. 38), allergies³⁹ and autoimmunity⁴⁰. Arrays are composed of known antigens, including intact antigenic particles, proteins, lipids, carbohydrates, linear peptides and constrained peptides in which disulphide bonds between cysteine residues provide secondary structure to the peptide^{40,41}. By 'printing' all such antigens on the same array, it is possible to gain valuable autoantibody-profiling data with a simple assay. Arrays have been constructed and validated for over a dozen autoimmune diseases, including connective-tissue diseases (such as SLE, scleroderma and myositis), primary biliary cirrhosis, experimental autoimmune encephalomyelitis and multiple sclerosis, rheumatoid arthritis, diabetes⁴², Crohn's disease, and sclerosing cholangitis^{40,41,43}. Such autoimmune-disease-specific arrays include self-antigens, viral proteins and peptides, and bacterial antigens with complex carbohydrates and recombinant proteins, such as flagellin⁴¹.

Arrays of antigens have been used to design and guide development of antigen-specific DNA vaccines for the treatment of multiple sclerosis and HIV infection³⁸. More specifically, arrays of autoantigens or viral antigens are printed and probed using serum derived from animal models of multiple sclerosis (mice with experimental allergic encephalomyelitis) or HIV (macaques with a simian version of HIV), then the antibody response is compared before and after a therapeutic intervention with a DNA vaccine encoding autoantigens or viral proteins, respectively. This line of experimentation has demonstrated that serum antibody epitope spreading is

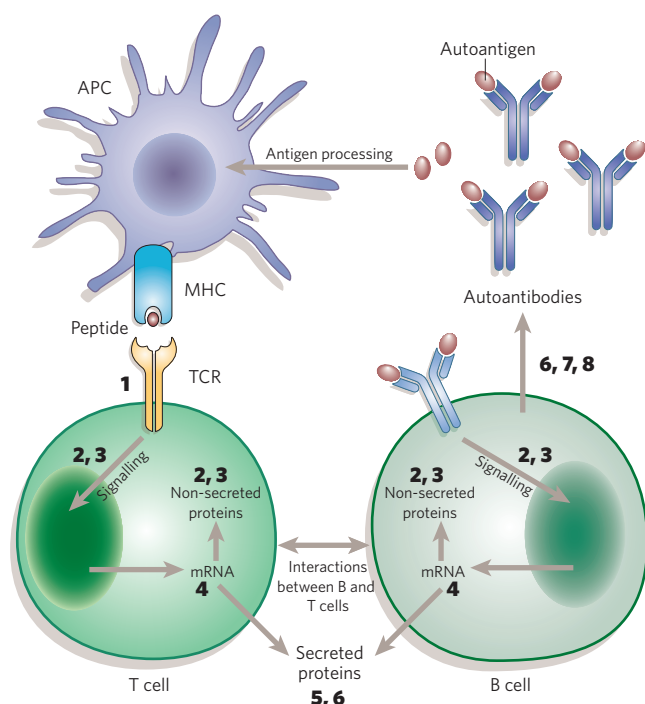


Figure 2 | Target sites of genomic and proteomic technologies. Discrete components of the autoimmune process from the presentation of MHC–peptide complexes to T cells to the production of autoantibodies by plasma cells can now be analysed in a multiplex fashion using various genomic and proteomic technologies. The numbers in the figure indicate the technologies that have demonstrated potential in the analysis of the corresponding components. (1) Peptide–MHC tetramer arrays. (2) Reverse-phase protein microarrays. (3) Multiparameter flow cytometry for intracellular antigens. (4) cDNA and oligonucleotide microarrays. (5) Antibody microarrays. (6) Bead-based multiplex assays. (7) Autoantigen microarrays. (8) Whole-proteome microarrays. Both mass spectrometry and 2D gel electrophoresis can be used to analyse complex mixtures of proteins and/or peptides.

altered as a result of therapy^{38,43}. This technique can be applied to any autoimmune disease in which one or more candidate target antigens have been identified. Large-scale arrays of recombinant proteins can also be produced that may allow the discovery of novel, unidentified autoantigens⁴⁴. Whether serum autoantibody profiles change in humans in response to therapeutic interventions remains to be studied.

Protein arrays for analysing T-cell receptors

Peptide–major histocompatibility complex (MHC) tetramer arrays have recently been developed and partially validated for the study of antigen-specific T cells⁴⁵. Individual peptide–MHC tetramer molecules are spotted onto the surface of glass microscope slides at specific positions before being probed with living T cells, which bind to individual tetramer spots and can be quantified and further studied by analysing calcium flux and cytokine secretion. Such assays are required for the detection of immune responses induced by antigen-specific therapies with peptides and proteins and by DNA vaccines (see the review by Feldmann and Steinman in this issue, page 612).

Antibody microarrays

In capture-antibody microarrays, immobilized antibodies trap their specific analytes from a sample solution (for example, serum, synovial fluid, culture supernatant or cell lysates). The bound antigens can then be detected through a direct protein-labelling approach (that is, fluorophores are covalently attached to the antigen) or a sandwich immunoassay approach (that is, a pair of antibodies recognizing two non-overlapping regions of the antigen, where one of them is fluores-

cently tagged). The intensity of fluorescence corresponds to the concentration of the antigen in the sample. Excellent reviews of this technology have recently been published, and so our discussion here is limited to its relevance to the study of autoimmunity⁴⁶.

Antibody arrays have been used to analyse intracellular protein levels and even their phosphorylation events, but successful application has largely been limited to the analysis of soluble cytokines and chemokines in culture supernatant and serum samples⁴⁷. Dysregulation of cytokine signalling is believed to play a crucial role in the initiation and maintenance of autoimmune diseases. For example, several lines of evidence support the notion that interferon- α is central to the development of SLE^{16,17}. Antibody microarrays are particularly suited to revealing unsuspected roles of cytokines in the development of autoimmune diseases. A recent study used antibody microarrays to simultaneously analyse the concentration of 78 cytokines, growth factors and soluble receptors in serum samples derived from patients with Crohn's disease and ulcerative colitis⁴⁸. Four cytokines were elevated in patients in clinical remission compared with patients with active disease. Among them was transforming growth factor- β (TGF- β), a cytokine that inhibits inflammatory activity and enhances regulatory T-cell functions. Although further studies are required to establish the roles of these cytokines, this is the first successful demonstration of the use of antibody microarrays in the study of autoimmune diseases.

Lysate arrays

One of the most promising techniques for studying blood cells and tissues targeted by the autoimmune response is the reverse-phase protein (RPP) lysate microarray platform⁴⁹. Lysates prepared from cells are deposited on slides and probed using antibodies of known specificity. Lysates can be prepared from tissue-culture cells, tissue-infiltrating autoreactive lymphocytes, diseased-tissue cells or blood cells and can be stimulated *in vitro* with agents, such as antigens, cytokines, drugs or antibodies that crosslink cellular receptors⁵⁰. Antibodies used to probe RPP microarrays include those that recognize housekeeping proteins, inducible factors, signalling proteins, cell-cycle regulatory proteins, apoptosis-related proteins and phosphorylation motifs present in signalling molecules. This approach has been highly successful in characterizing the activation state of tumour cells⁴⁷ and more recently regulatory T cells^{50,51}. By combining RPP microarrays with laser-capture microdissection (a technique that allows the specific capture and study of individual cells from tissue sections or histological specimens fixed to microscope slides), it should be possible to study rare cells (such as the lymphocytes that infiltrate diseased tissue), dendritic cells and autoimmune target cells, such as glomerular cells in SLE nephritis, neuronal cells in multiple sclerosis or β -cells or islet-infiltrating T cells in insulinitis.

Assays based on flow cytometry

Fluorescence-activated cell sorting (FACS) has revolutionized the study of immunology and has recently been adapted for the study of intracellular signalling pathways. In this technique, cell populations are first identified using monoclonal antibodies that are conjugated to spectrally resolvable fluorophores and are specific for cell-surface proteins. Cells are then fixed and permeabilized before staining with antibodies that recognize intracellular proteins, including cytokines, chemokines, structural proteins, apoptosis-related proteins or signalling molecules, such as kinases⁵². Unlike any of the above technologies, intracellular FACS analysis with multiple fluorochromes allows the characteristics of a single cell to be studied, rather than a mixture of cells. Most recently, this approach has been extended to incorporate the use of phospho-specific antibodies⁵² and has been validated for characterizing subsets of patients with acute myeloid leukaemia⁵³. Combining this approach with RPP microarrays and multiplexed cytokine assays holds unlimited promise for characterizing autoimmune diseases.

Multiplexed bead-based flow-cytometry assays represent an active

area of development. In these assays, differentially identifiable beads are coated with proteins, including antibodies and autoantigens and then identified using a flow cytometer. Such bead-based assays have been adapted for measuring cytokine concentrations in serum or culture supernatants and autoantibodies in serum samples derived from patients with autoimmune disease⁴⁶. Advances in instrumentation and bead chemistries will probably make this approach very valuable for the study of autoimmunity.

Conclusion and future directions

The expanding use of genomic and proteomic approaches in the analysis of autoimmune disease and therapy has identified numerous areas for improvement and further development. Although DNA microarray studies have led to important advances in the study of autoimmunity, their use is limited owing to the problem of physiologically important splice variants that exist in individual cell populations. In partial response to this problem, several commercial entities are developing new technologies that use exons or single-nucleotide arrays for tiling, allowing additional information on alternative splice variants and polymorphisms. Unfortunately, these new data require additional sophistication in bioinformatics to analyse the increasingly complex data. An additional problem lies in the fact that a large proportion of mRNAs are not translated into proteins, despite being upregulated at the level of transcription.

The proteomics field remains in its infancy and is limited largely by current technology and the dearth of high-quality reagents (specifically monoclonal antibodies) and informatics tools. Unbiased proteomic techniques have largely been disappointing for studying autoimmune diseases, while array and FACS platforms are limited because it is impossible to know which important aspects will be missed owing to the limitations of current arrays. Definition of the interactome is sadly lacking in most mammalian systems. Computing algorithms and two-hybrid systems of interaction need to be generated to allow disparate datasets to be studied simultaneously — such as transcript profiling, protein array and interaction datasets, and FACS. This is no small task. Finally, standards for proteomic work and public datasets need to be developed. Taken together, a multidisciplinary approach to the study of autoimmunity will be required in the coming decade, an approach that combines the skills of biologists, clinicians, engineers and bioinformaticians. ■

- Ermann, J. & Fathman, C. Autoimmune diseases: genes, bugs and failed regulation. *Nature Immunol.* **2**, 759–766 (2002).
- Buckner, J. & Nepom, G. Genetics of rheumatoid arthritis: is there a scientific explanation for the human leukocyte antigen association? *Curr. Opin. Rheumatol.* **14**, 254–259 (2002).
- Ye, S., Usher, D. & Zhang, L. Gene expression profiling of human diseases by serial analysis of gene expression. *J. Biomed. Sci.* **5**, 384–394 (2002).
- Moser, K. et al. The use of microarrays to study autoimmunity. *J. Invest. Dermatol. Symp. Proc.* **1**, 18–22 (2004).
- Quackenbush, J. Computational analysis of microarray data. *Nature Rev. Genet.* **2**, 418–427 (2001).
- Tusher, V., Tibshirani, R. & Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA* **98**, 5116–5121 (2001).
- Eisen, M., Spellman, P., Brown, P. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA* **95**, 14863–14868 (1998).
- Iyer, V. et al. The transcriptional program in the response of human fibroblasts to serum. *Science* **283**, 83–87 (1999).
- Johnson, J. et al. Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science* **302**, 2141–2144 (2003).
- Ng, B. et al. Increased noncanonical splicing of autoantigen transcripts provides the structural basis for expression of untolerized epitopes. *J. Allergy Clin. Immunol.* **6**, 1463–1470 (2004).
- Soares, L. et al. Two isoforms of *otubain 1* regulate T cell anergy via GRAIL. *Nature Immunol.* **1**, 45–54 (2004).
- Castle, J. et al. Optimization of oligonucleotide arrays and RNA amplification protocols for analysis of transcript structure and alternative splicing. *Genome Biol.* **10**, R66 (2003).
- Lock, C. et al. Gene microarray analysis of multiple sclerosis lesions yields new targets validated in autoimmune encephalomyelitis. *Nature Med.* **8**, 500–508 (2002).
- Chabas, D. et al. The influence of the proinflammatory cytokine, osteopontin, on autoimmune demyelinating disease. *Science* **294**, 1731–1735 (2001).
- Karpui, M. et al. Prolonged survival and decreased abnormal movements in transgenic model of Huntington disease, with administration of the transglutaminase inhibitor cystamine. *Nature Med.* **8**, 143–149 (2002).
- Baechler, E. et al. Interferon-inducible gene expression signature in peripheral blood cells of patients with severe lupus. *Proc. Natl Acad. Sci. USA* **100**, 2610–2615 (2003).
- Crow, M. & Wohlgemuth, J. Microarray analysis of gene expression in lupus. *Arthritis Res. Ther.* **6**, 279–288 (2003).
- Rus, V. et al. Expression of cytokine and chemokine related genes in peripheral blood mononuclear cells from lupus patients by cDNA array. *Clin. Immunol.* **102**, 283–290 (2002).
- Aune, T. et al. Co-localization of differentially expressed genes and shared susceptibility loci in human autoimmunity. *Genet. Epidemiol.* **2**, 162–172 (2004).
- Robinson, W., Steinman, L. & Utz, P. Proteomics technologies for the study of autoimmune disease. *Arthritis Rheum.* **46**, 885–893 (2002).
- Golub, T. et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**, 531–537 (1999).
- Alizadeh, A. et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403**, 503–511 (2000).
- Ito, T. et al. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl Acad. Sci. USA* **98**, 4569–4574 (2001).
- Li, S. et al. Map of the interactome network of the metazoan *C. elegans*. *Science* **303**, 540–543 (2004).
- Quackenbush, J. Microarrays—guilt by association. *Science* **302**, 240–241 (2003).
- Marcotte, E., Pellegrini, M., Thompson, M., Yeates, T. & Eisenberg, D. A combined algorithm for genome-wide prediction of protein function. *Nature* **402**, 83–86 (1999).
- Rain, J. et al. The protein–protein interaction map of *Helicobacter pylori*. *Nature* **409**, 211–215 (2001).
- Walhout, A. et al. Protein interaction mapping in *C. elegans* using proteins involved in vulval development. *Science* **287**, 116–122 (2000).
- Gygi, S., Corthals, G., Zhang, Y., Rochon, Y. & Aebersold, R. Evaluation of two-dimensional gel electrophoresis-based proteome analysis technology. *Proc. Natl Acad. Sci. USA* **97**, 9390–9395 (2000).
- Coomes, K. R., Morris, J. S., Hu, J., Edmonson, S. R. & Baggerly, K. A. Serum proteomics profiling — a young technology begins to mature. *Nature Biotechnol.* **23**, 291–292.
- Liao, H. et al. Use of mass spectrometry to identify protein biomarkers of disease severity in the synovial fluid and serum of patients with rheumatoid arthritis. *Arthritis Rheum.* **50**, 3792–3803 (2004).
- Dumont, D., Noben, J., Raus, J., Stinissen, P. & Robben, J. Proteomic analysis of cerebrospinal fluid from multiple sclerosis patients. *Proteomics* **4**, 2117–2124 (2004).
- Drynda, S. et al. Proteome analysis reveals disease-associated marker proteins to differentiate RA patients from other inflammatory joint diseases with the potential to monitor anti-TNF α therapy. *Pathol. Res. Pract.* **200**, 165–171 (2004).
- Stone, J. et al. A serum proteomic approach to gauging the state of remission in Wegener's granulomatosis. *Arthritis Rheum.* **52**, 902–910 (2005).
- Lorenz, P., Ruschpler, P., Koczan, D., Stiehl, P. & Thiesen, H. From transcriptome to proteome: differentially expressed proteins identified in synovial tissue of patients suffering from rheumatoid arthritis and osteoarthritis by an initial screen with a panel of 791 antibodies. *Proteomics* **3**, 991–1002 (2003).
- Macbeath, G. Protein microarrays and proteomics. *Nature Genet.* **32**, 526–532 (2002).
- Wilson, D. & Nock, S. Recent developments in protein microarray technology. *Angew. Chem.* **42**, 494–500 (2003).
- de Vegvar, H. et al. Microarray profiling of antibody responses against Simian-Human Immunodeficiency Virus: Postchallenge convergence of reactivities independent of host histocompatibility type and vaccine regimen. *J. Virol.* **77**, 11125–11138 (2003).
- Wiltshire, S. et al. Detection of multiple allergen-specific IgEs on microarrays by immunosorbent with rolling circle amplification. *Clin. Chem.* **46**, 1990–1993 (2000).
- Robinson, W. et al. Autoantigen microarrays for multiplex characterization of autoantibody responses. *Nature Med.* **8**, 295–301 (2002).
- Wang, D., Liu, S., Trummer, B., Deng, C. & Wang, A. Carbohydrate microarrays for the recognition of cross-reactive molecular markers of microbes and host cells. *Nature Biotechnol.* **20**, 275–281 (2002).
- Quintana, F. et al. Functional immunomics: microarray analysis of IgG autoantibody repertoires predicts the future response of mice to induced diabetes. *Proc. Natl Acad. Sci. USA* **101**, 14615–14621 (2004).
- Robinson, W. et al. Protein microarrays guide tolerizing DNA vaccine treatment of autoimmune encephalomyelitis. *Nature Biotechnol.* **21**, 1033–1039 (2003).
- Michaud, G. et al. Analysing antibody specificity with whole proteome microarrays. *Nature Biotechnol.* **21**, 1509–1512 (2003).
- Soen, Y., Chen, D., Kraft, D., Davis, M. & Brown, P. Detection and characterization of cellular immune responses using peptide-MHC microarrays. *PLoS Biol.* **1**, E65 (2003).
- Utz, P. Protein arrays for studying blood cells and their secreted products. *Immunol. Rev.* **204**, 264–282 (2005).
- Gilburd, B. et al. Autoantibodies profile in the sera of patients with Sjögren's syndrome: the ANA evaluation—a homogeneous, multiplexed system. *Clin. Dev. Immunol.* **11**, 53–56 (2004).
- Kader, H. A. et al. Protein microarray analysis of disease activity in pediatric inflammatory bowel disease demonstrates elevated serum PLGF, IL-7, TGF- β 1, and IL-12p40 levels in Crohn's disease and ulcerative colitis patients in remission versus active disease. *Am. J. Gastroenterol.* **100**, 414–423.
- Espina, V. et al. Protein microarrays: molecular profiling technologies for clinical specimens. *Proteomics* **3**, 2091–2100 (2003).
- Chan, S., Ermann, J., Su, L., Fathman, C. & Utz, P. Protein microarrays for multiplex analysis of signaling pathways. *Nature Med.* **10**, 1390–1396 (2004).
- Su, L. et al. Murine CD4⁺CD25⁺ regulatory T cells fail to undergo chromatin remodeling across the proximal promoter region of the IL-2 gene. *J. Immunol.* **173**, 4994–5001 (2004).
- Perez, O. & Nolan, G. Simultaneous measurement of multiple active kinase states using polychromatic flow cytometry. *Nature Biotechnol.* **20**, 155–162 (2002).

53. Irish, J. *et al.* Single cell profiling of potentiated phospho-protein networks in cancer cells. *Cell* **118**, 217–228 (2004).
54. Petricoin, E. *et al.* Use of proteomic patterns in serum to identify ovarian cancer. *Lancet* **359**, 572–577 (2002).
55. Zhu, H. *et al.* Global analysis of protein activities using proteome chips. *Science* **293**, 2101–2105 (2001).
56. Joos, T. *et al.* A microarray enzyme-linked immunosorbent assay for autoimmune diagnostics. *Electrophoresis* **21**, 2641–2650 (2000).
57. Schweitzer, B. *et al.* Multiplexed protein profiling on microarrays by rolling circle amplification. *Nature Biotechnol.* **20**, 359–365 (2002).
58. Liotta, L. *et al.* Protein microarrays: meeting analytical challenges for clinical applications. *Cancer Cell* **3**, 317–325 (2003).

Acknowledgements The authors thank members of the Utz and Fathman labs for insights and scientific creativity. CGF is supported by the JDRFI, ADA, UCSF and the NIH. Steven Chan is supported by the Stanford MSTP, Bio-X and the Floren Family Foundation. Luis Soares is supported by the NIH. PJU is the recipient of a Donald E. and Delia B. Baxter Foundation Career Development Award and was supported by the Dana Foundation, the Floren Family Foundation, the Stanford Program in Molecular and Genetic Medicine (PMGM), NIH Grants and NHLBI Proteomics Consortium.

Competing interests statement The authors declare competing financial interests: details accompany the paper on www.nature.com/nature.

Design of effective immunotherapy for human autoimmunity

Marc Feldmann¹ & Lawrence Steinman²

A better understanding of the molecules involved in immune responses has identified many potential targets for the treatment of autoimmune diseases. But although successful therapies have been found for immune disorders in animal studies, few have passed the much harder test of treating human diseases. So far, non-antigen-specific approaches, such as the blocking of tumour-necrosis factor, are achieving some success but the same is not true for antigen-specific approaches. Future therapies will probably include both non-antigen-specific strategies that target cytokines (cell-cell signalling molecules) or block the molecules that stimulate immune responses, and antigen-specific therapies that induce tolerance to self antigens.

Immunotherapy is a type of treatment that uses immunological tools, such as monoclonal antibodies, receptor-immunoglobulin fusion proteins, vaccines and immune cells. Such therapeutic options have only been available in the past 10 to 15 years, owing to major advances in medical science and technology, but are now increasingly being used to tackle a wide spectrum of human diseases. The application of immunotherapy to autoimmune diseases is broadening our understanding of the human immune response, with responses to treatment providing unique insights into pathological mechanisms. The availability of effective immunosuppressive drugs¹ to ameliorate the immune-mediated rejection of transplants contrasts sharply with the paucity of drugs that successfully treat autoimmune diseases. This implies that whereas a transplant is a classic acute challenge to an otherwise normal immune system, chronic autoimmune diseases are somehow different.

The failure of most immunological approaches that are effective in animal models^{2,3} to modulate autoimmune disease in humans suggests that we do not understand many of the principles behind the pathogenic mechanisms of these diseases. We remain ignorant of what drives the chronicity of these conditions, which can last for decades, and of how we can normalize the immune and pro-inflammatory responses once they commence. The rate-limiting steps of the early immune response (such as the presentation of antigen by dendritic cells, the expansion of CD4⁺ helper T-cell populations and the induction of costimulatory-molecule expression) may not be rate limiting or critical for the chronic phase of the disease and the resultant tissue destruction, which often occur years after onset.

Human transplants, which often undergo chronic rejection¹ despite continuous immunosuppressive therapy and early success, have confirmed our lack of understanding of chronicity. Furthermore, results in acute animal models of autoimmunity are often not predictive for the treatment of chronic human immune disorders²⁻⁴. Because we do not understand the differences between the chronic and acute response, we cannot be sure which, if any, animal models of disease provide good reflections of the key processes that occur in human disease. A further complication for the transition from animal to human studies is the necessary preoccupation with safety in human immunotherapy, a relatively ignored issue in animal models.

Here, we highlight recent successes in immunotherapy, which is now benefiting almost a million patients with chronic diseases, such as rheumatoid arthritis and Crohn's disease, that are unresponsive to other treatments. We contrast the effectiveness of therapies aimed at inhibiting the non-antigen-specific pathways, such as cytokine and cell-trafficking pathways (components of innate immunity), with the comparative lack of success of therapies that interfere with the more complex and flexible features of antigen-specific adaptive immunity.

Targets for immunotherapy

The treatment of human autoimmune diseases often occurs years after the onset of the pathogenic process, and despite our increasing knowledge of the cellular and molecular processes involved in immunity, the most effective targets for immunotherapy in the chronic phase of the disease are not obvious. Targeting various critical molecules involved in pathological pathways has led to the modulation of disease in animal models (Fig. 1). Components of the pathological cascade that have received most attention are: factors involved in lymphocyte homing to target tissues; enzymes that are critical for the penetration of blood vessels and the extracellular matrix by immune cells; cytokines that mediate pathology within the tissues; various cell types that mediate the damage at the site of the disease, as well as these cells' antigen-specific adaptive receptors, including the T-cell receptor (TCR) and immunoglobulin; and other toxic mediators, such as complement components and nitric oxide (Fig. 1).

A widespread misconception is that every step of the immune or pro-inflammatory process is a potential therapeutic target. Regrettably, this is not the case. Because most therapeutics only have a partial inhibitory effect, only those molecules that are in short supply (and thus rate-limiting) are likely to be useful targets. Therefore, therapy that specifically targets most of the steps (which are non-rate-limiting) in the immune or pro-inflammatory process yields little benefit in ongoing (late, active) autoimmune disease in humans. So far, only therapies that target two rate-limiting steps — the cytokine tumour-necrosis factor (TNF; ref. 5) and the molecule involved in lymphocyte homing, $\alpha_4\beta_1$ integrin⁶ — have markedly ameliorated autoimmune disease progression; for example, in rheumatoid arthritis, inflamma-

¹Kennedy Institute of Rheumatology, Faculty of Medicine, Imperial College School of Medicine, ARC Building, 1 Aspenlea Road, London W6 8LH, UK (e-mail: m.feldmann@imperial.ac.uk).

²Department of Neurology and Neurological Sciences, Stanford University, Stanford, California 94305, USA (e-mail: Steinman@stanford.edu).

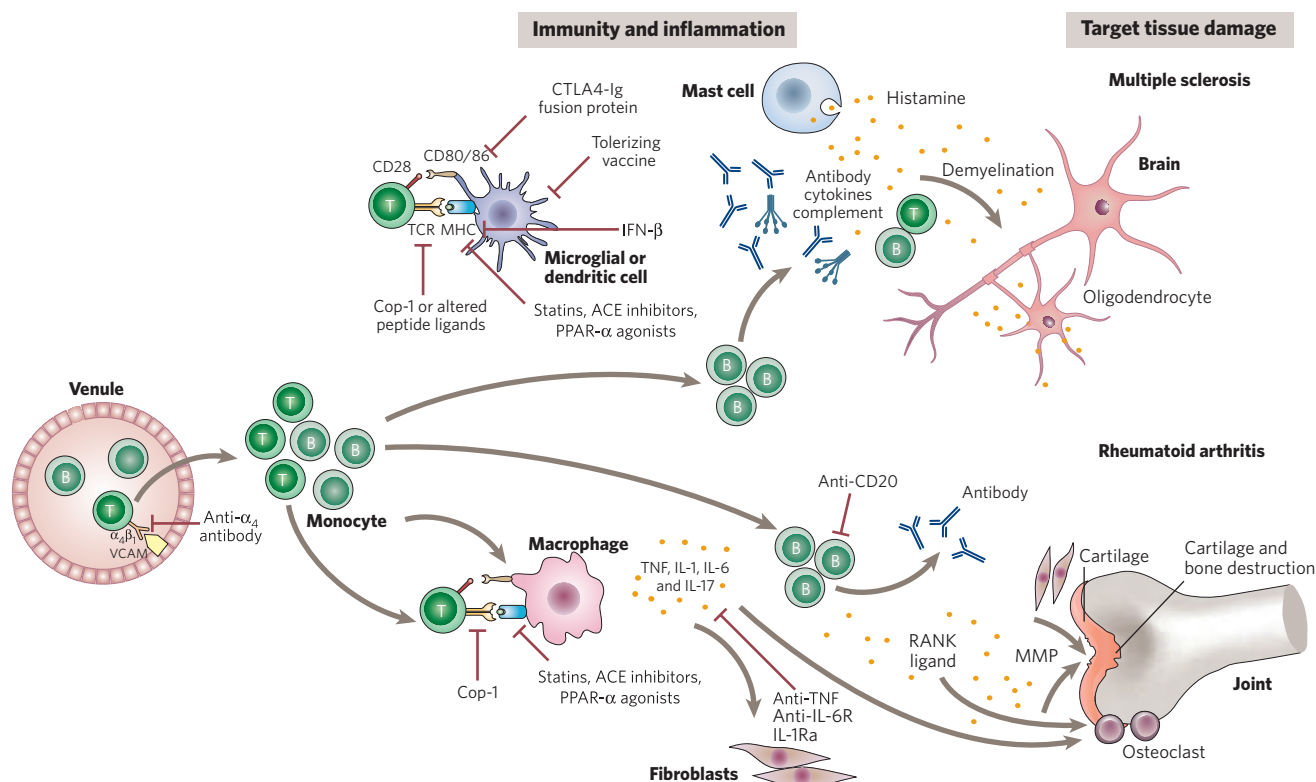


Figure 1 | Pathogenesis of multiple sclerosis and rheumatoid arthritis.

Activated T cells express $\alpha_4\beta_1$ -integrin, which binds to vascular cellular adhesion molecule (VCAM) on the surface of venules in inflamed tissues. This interaction allows the T cells to pass through the endothelial wall and penetrate the extracellular matrix. In multiple sclerosis (upper panel), the T cells re-encounter the cognate CNS antigen presented by MHC class II molecules on either microglial or dendritic cells. This interaction can be inhibited by glatiramer acetate (Cop-1) or altered peptide ligands. In addition, statins, angiotensin-converting enzyme (ACE) inhibitors, and PPAR- α agonists can all downregulate the inducible expression of MHC class II molecules. Similarly, cytokines such as interferon- β (IFN- β) downregulate MHC class II molecules and interfere with diapedesis of cells (the movement of cells through the endothelial wall) by downregulating metalloproteases. CD28 and CD80/86 interactions can be blocked by the CTLA4-Ig fusion protein. Tolerizing vaccines promote tolerance processes which occur when the T cell/dendritic

cell interaction is not optimal. B cells and mast cells are also recruited into the inflammatory infiltrate. Antibody plus complement can produce 'membrane attack' complexes, which can damage the oligodendrocytes and underlying axon. Osteopontin is expressed on the surface of oligodendroglial cells and neurons during active disease, and is pivotal in the disease progression. In rheumatoid arthritis, T cells and macrophages that have entered the synovium from inflamed venules produce cytokines, especially TNF, IL-1, IL-6 and IL-17, which mediate damage to the synovium. This damage can be blocked by anti-TNF antibody, IL-1 receptor antagonist (IL-1Ra), and anti-IL-6 receptor (IL-6R) antibody. RANK ligand is the main signal for activating osteoclasts in cartilage, which mediate bone destruction. Anti-TNF antibodies reduce the migration of lymphocytes from the blood to the synovium, and also prevent bone loss by blocking the destructive effects of IL-1, IL-6 and TNF. Anti-CD20 kills B cells but not plasma cells; fibroblasts make most of the tissue-destructive metalloproteinases (MMPs).

tory bowel disease, ankylosing spondylitis, psoriasis and multiple sclerosis.

These particular key molecules and the processes they control can be referred to as 'tipping points'⁷. In epidemiology, a tipping point is defined as the moment when epidemics qualitatively change, reach a critical mass and have major repercussions. This concept is valuable in autoimmune diseases because many cellular and molecular processes contribute to tipping the balance towards the disease state, and therefore are potential therapeutic targets. But although targeting these tipping points may provide significant benefit, in terms of treating autoimmune disease, blocking these critical physiological molecules could also negate their beneficial roles in generating protective immune responses, and therefore could lead to an increased risk of infection. For example, despite the enormous success in treating multiple sclerosis by blocking $\alpha_4\beta_1$ integrin, this treatment was recently voluntarily withdrawn because of the development of a fatal untreatable infection. So tipping points are physiological processes that are key to maintaining both health and disease.

The targeting of TNF (ref. 8) or $\alpha_4\beta_1$ (refs 6, 9, 10) has remarkable effects on several autoimmune diseases, including rheumatoid arthritis, inflammatory bowel disease, ankylosing spondylitis, psoriasis and multiple sclerosis. These molecules can therefore be considered as true tipping points in the pathophysiology of autoimmune disease. But

Box 1 | Using commonly used drugs to treat autoimmunity

Recently, familiar oral medications, such as statins and angiotensin blockers, widely used for other disease conditions such as hypercholesterolaemia, hypertriglyceridaemia, allergy and hypertension, have been shown to inhibit some of the biochemical reactions that occur in autoimmune inflammation (Fig. 1). These drugs have shown promise in pre-clinical models of autoimmunity, as well as in early-stage clinical trials¹¹. Even if they are not optimal therapies on their own, they are clearly pointing towards key alternative pathways, and may prove to be effective when used in synergy with other approaches.

Interestingly, the statins, which block the activity of the enzyme 3-hydroxy-3-methylglutaryl-coenzyme A (HMG-CoA) reductase, and thus reduce levels of cholesterol, also inhibit the appearance of inducible MHC class II molecules⁷². The statins are remarkably potent in reversing disease in animal models, inducing shifts from the production of T_H1 -type pro-inflammatory cytokines by autoaggressive T cells to T_H2 -type cytokines⁷⁴. Initial trials administering statins to multiple sclerosis and rheumatoid arthritis patients show moderate efficacy^{71,75}. As with statins, peroxisome proliferator-activated receptor- α (PPAR- α) agonists — drugs used in type II diabetes — which regulate the activation of adipocytes and macrophages, also induce a shift in cytokine production from the T_H1 to T_H2 type⁷⁶. Initial experiments suggest angiotensin blockers do the same⁷⁷. The efficacy of the statins, PPAR- α agonists and angiotensin blockers may result from their ability to alter a number of pathological processes in the immune cascade.

because the benefit seen here is achieved by interfering with processes that are involved in both host defence and autoimmune pathology, the overall benefit:risk ratio is inherently difficult to predict.

In the pharmaceutical industry, drugs in only about 5% of the 'small-molecule' drug projects end up as approved therapeutics; most drop out because of problems, usually toxicity. Hence, existing drugs (which are relatively safe) with new potential uses present a wonderful opportunity. Recently, familiar oral medications, such as statins, which are widely used for other disease conditions, have been shown to be effective in animal models of autoimmunity and early-stage clinical trials in patients with multiple sclerosis and rheumatoid arthritis¹¹ (Box 1).

Non-antigen-specific approaches

T-cell populations and antigen-presenting cells

Despite preventing disease (such as arthritis and experimental autoimmune encephalomyelitis, EAE), to an impressive extent in animal models, anti-CD4-antibody therapy, with either lytic or non-lytic monoclonal antibodies, has not successfully treated human rheumatoid arthritis¹², psoriasis or multiple sclerosis¹³. However, the limited scope for experimentation in humans during clinical trials may mean that inappropriate antibodies or dose regimes have been used. Alternatively, failure to prevent disease might have been caused by the anti-CD4 antibody also inhibiting regulatory T cells that express CD4. By contrast, encouraging results have been reported from both animal models¹⁴ and early clinical studies¹⁵ using a mutated, less activating form of anti-CD3 antibody. The use of this antibody avoids the acute cytokine release — that causes a range of problems from malaise to hypotensive shock¹⁶ — induced by non-mutated anti-CD3 antibody.

There is a growing consensus that antigen-presenting cells (APCs) are important rate-limiting cells for inducing immune responses¹⁷: a leading hypothesis is that inducible major histocompatibility complex (MHC) class II molecule expression is induced inappropriately on APCs at the site of autoimmune disease¹⁸ (Fig. 1). Consistent with this, in many animal models of autoimmune disease, antibodies specific for MHC class II molecules reduce disease. But because the antibodies caused unexpected toxicity when tested in monkeys¹⁹, this has not yet been tested in humans.

Effective antigen presentation and activation of T cells requires not only TCR recognition of MHC molecules complexed with a peptide, but also various ligand–receptor costimulatory interactions at the 'immune synapse' — the point of interaction between a T cell and an APC. Most important among these costimulatory interactions are CD28 molecules recognizing CD80 or CD86 molecules¹¹ (Fig. 2). Therapy using a cytotoxic T-lymphocyte antigen 4 (CTLA4)–immunoglobulin fusion protein, which blocks interactions with CD28, is effective in randomized, double-blind clinical trials in patients with psoriasis and rheumatoid arthritis²⁰, suggesting that even in late-stage disease, signals mediated by costimulatory molecules expressed by APCs are required. Blocking other molecules that are involved in activating the immune system may also be useful therapeutically. Unfortunately, despite promising results in experimental studies, the administration of an antibody specific for the T-cell-expressed costimulatory molecule CD40 ligand was toxic in humans, causing a number of deaths from thrombosis. The blocking of costimulatory molecules that are expressed only after antigen activation of T cells, such as OX40, may be efficacious and safer²¹, as this would not block uninvolved T cells.

Regulatory T cells and B cells

Several regulatory subsets of T cells have been defined in recent years, and attention is now turning to their use for therapy. This is because defects in such regulatory subsets (in particular, the CD4⁺CD25⁺ regulatory T-cell subset) may be important in enabling autoimmune diseases to become established^{22,23} (see review by Kronenberg and Rudensky in this issue, page 598).

Given the ubiquity of autoantibodies in autoimmune diseases, it was assumed that the antibody-producing cells — plasma cells and B cells

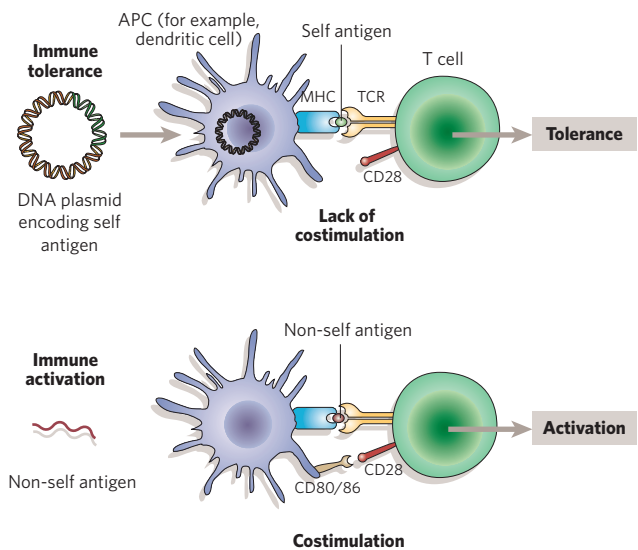


Figure 2 | Generating immune tolerance by using 'tolerizing' DNA vaccines.

A DNA plasmid encoding a self antigen is transcribed and translated in a dendritic cell, but its expression does not stimulate the innate immune system enough to upregulate costimulatory molecules. A further reduction in costimulation is caused by the removal of CpG motifs in the plasmid. The presentation of self antigen by APCs without adequate costimulation leads to anergy or tolerance of T cells, because of the lack of interaction between CD28 with CD80 or CD86 (refs 72, 73). In contrast, conventional immunization, with a foreign antigen, leads to effective presentation of antigen in the MHC molecules with adequate costimulation, and leads to productive cytokine cascades and gene activation.

— would be a good target for therapy. However, this assumption has only recently been confirmed: lytic anti-CD20 antibody (rituximab; Rituxan), which lyses B cells, effectively treated rheumatoid arthritis and systemic lupus erythematosus²⁴, although extensive comedication of subjects in these trials makes the data difficult to interpret.

Cytokines

Cytokines are short-range protein mediators with a wide range of actions. They are important in all biological processes²⁵, including T-cell growth (IL (interleukin)-2, IL-4, IL-7, IL-15 and IL-21), inflammation (TNF, IL-1, IL-6 and IFN (interferon)- γ) as well as the inhibition of inflammation (IL-10, transforming growth factor- β (TGF- β) and IL-4). As extracellular molecules, they are accessible to 'biologics' — protein therapeutics such as antibodies or soluble receptors.

The relative potency of cytokines that induce multiple biological effects is compatible with a rate-limiting, 'catalytic' role, and therefore they are potential therapeutic targets. A major problem in establishing which ones may be targets lies in the considerable overlap (redundancy) in their biological properties. Thus, IL-1, TNF, IL-6 and granulocyte–macrophage colony-stimulating factor (GM-CSF) have more than 80% overlap in function, when tested *in vitro*. So, which ones are likely to be therapeutic targets in which diseases? Insights into this problem have come both from *in vivo* experiments using animal models and from clinical studies²⁶.

In contrast to the limited success of treatment with cytokines (see below), blockade of cytokines is the success story of the current era of molecular therapy in autoimmunity, which is based on scientific analysis of disease mechanisms⁸. Research using joint tissue from patients with rheumatoid arthritis suggested the importance of TNF in the disease pathogenesis²⁷. The existence of TNF-inhibiting biologics (originally generated to treat sepsis syndrome) made it possible to perform a successful proof-of-principle clinical trial in 1992 (ref. 28) with the anti-TNF monoclonal antibody infliximab. This culminated

in the approval from 1998/1999 of a set of therapeutic biologicals: anti-TNF monoclonal antibodies (infliximab²⁹ and adalimumab³⁰) and the TNF-receptor (TNFR) fusion protein (etanercept³¹; Enbrel). TNF blockade has demonstrated that biologicals can be used in the long term, and extensively: about a million patients have been treated with anti-TNF biologicals so far, and some for over seven years.

Much has been learnt from the use of anti-TNF biologicals; for example, the importance of finding the right therapeutic target. TNF is the body's fire alarm⁵. It initiates the defence response to local injury, recruits leukocytes, and initiates a whole series of events that are important in health and in many diseases. Hence its blockade is useful in treating many diseases. The mechanism-of-action studies (see Box 2) have provided several insights into the pathogenesis of targeted diseases³², especially rheumatoid arthritis.

IL-6 is another useful target, with clinical-trial success in rheumatoid arthritis showing comparable efficacy to TNF blockade³³. However, the clinical benefits of IL-6 blockade occur more slowly than with TNF blockade, as predicted from *in vitro* studies that revealed a TNF-dependent cytokine cascade^{27,34}, where TNF drives the production of multiple pro-inflammatory cytokines. Success has also come from IL-1 blockade using the IL-1-receptor antagonist anakinra, which is approved for the treatment of rheumatoid arthritis³⁵. And promising results have been seen in the treatment of rheumatoid arthritis with anti-IL-15 antibody³⁶. High mobility group 1 (HMGB1), a stimulator of inflammatory responses, is another promising target for arthritis and sepsis³⁷. Finally, blocking the receptor activator of nuclear factor κ B ligand (RANKL), the main activator of osteoclasts, is a promising approach for reducing bone destruction, such as that seen in rheumatoid arthritis³⁸.

Cell recruitment: chemokines and adhesion molecules

The small-protein chemotactic cytokines (chemokines) have several properties that make them favoured targets in the pharmaceutical industry³⁹: they are extracellular, and so accessible to biologicals; and

they bind to seven-transmembrane receptors that can be blocked by small-molecular-mass chemicals. Most importantly, chemokines are mediators of cell migration. Because chronic inflammatory diseases depend on the recruitment of inflammatory cells to the inflamed site, any approach that reduces the number of inflammatory cells in the site of disease may be of benefit, be it through chemokine or adhesion-molecule blockade. However, like cytokines, there are numerous chemokines (more than 40) with redundant properties, so it is not clear which ones are the most relevant in which disease.

Immune surveillance is accomplished by highly mobile leukocytes that are primed to fight microbes anywhere in our bodies. Organ-specific autoimmunity may result when autoreactive lymphocytes enter an inflamed site, initiating multiple events^{8,18}. Lymphocyte migration depends on highly specific 'adhesion' molecules expressed by T cells that bind to receptors induced on endothelial cells⁴⁰. These adhesion molecules and their receptors have domains in the extracellular space, and so can be targeted with monoclonal antibodies. Because the key homing molecules — integrins and selectins — display a high degree of diversity, a particular integrin molecule or selectin molecule is critical for entry to a particular anatomical site, and blocking that molecule might abolish pathological homing to that site, leaving lymphocytes free to move elsewhere.

Initial studies in animal models of multiple sclerosis (EAE) indicated that the critical homing molecule to the inflamed central nervous system (CNS) is $\alpha_4\beta_1$ integrin⁹: anti- $\alpha_4\beta_1$ antibody blocked the entry of lymphocytes into the brain and abrogated the clinical paralysis associated with EAE. This approach also proved successful in patients with multiple sclerosis: a phase III trial of a humanized $\alpha_4\beta_1$ -specific monoclonal antibody natalizumab (Tysabri) reduced clinical relapses by 66% over the next year, leading to Food and Drug Administration (FDA) approval of the drug⁶. Encouraging results were seen with the same antibody in the treatment of inflammatory bowel disease¹⁰.

However, the blockade of $\alpha_4\beta_1$ integrin is not specific. It interferes with lymphocyte homing in general, and therefore raises the risk of opportunistic infections¹¹. Recently, sales of natalizumab were withdrawn, after two patients taking it in combination with IFN- β 1a (Avonex) developed progressive fatal multifocal leukoencephalopathy, an untreatable viral infection (<http://www.fda.gov/cder/drug/infopage/natalizumab/default.htm>). Blocking lymphocyte mobility with these two drugs, and blocking lymphocyte entry to the brain, may have caused this unusual infection, caused by the ubiquitous JC virus, the activation of which is most commonly seen in severely immunocompromised individuals.

Antigen-specific approaches

The adaptive autoimmune response becomes more complex as disease progresses, owing to the generation of T-cell reactivity and antibodies to other local molecules — a concept known as epitope spreading⁴¹. Thus, in the chronic stage of the disease, the adaptive immune response targets several different molecules at the anatomical site of the disease.

In the 1970s, a random copolymer of the amino acids glutamate, tyrosine, alanine and lysine (copolymer 1 or Cop-1), now termed glatiramer acetate or copaxone, was designed to mimic the composition of myelin basic protein (MBP) — a major target of autoimmune responses in multiple sclerosis. The administration of glatiramer acetate ameliorated EAE, and is now an approved drug for multiple sclerosis⁴²: daily injection of glatiramer acetate reduces disease relapse by 30%, and induces a T helper 2 (T_H2)-type response to myelin antigens. This is desirable because T_H1 -type responses to myelin proteins are pathogenic. However, T_H2 -type responses are associated with allergic reactions, and about 10% of individuals taking glatiramer acetate develop allergic reactions.

An altered peptide ligand (APL) of MBP-derived peptide 83–99 was constructed by mutating the amino acids that form the main contact sites with the TCR on disease-causing T cells⁴³. The administration of the

Box 2 | Anti-TNF therapy of rheumatoid arthritis

Mechanism of action

- Reduction in pro-inflammatory cytokine cascade, including reduction of IL-6, IL-1, GM-CSF and vascular endothelial growth factor (VEGF).
- Reduction in leukocyte trafficking owing to decreased expression of adhesion molecules and chemokines.
- Reduction in tissue-destructive enzymes, such as matrix metalloproteinases (MMPs), but levels of tissue inhibitor of MMPs are maintained.
- Reduction in angiogenesis through reduced VEGF production.
- Normalization of abnormal haematology: haemoglobin restored, platelets and fibrinogen reduced.

Clinical benefits

- Reduction of symptoms including pain, stiffness and lethargy.
- Reduction in signs of active disease including tenderness and joint swelling.
- Reduction in cartilage and bone damage.
- Induction of tissue repair.

Potential side effects

- Increased risk of infection due to reduced cytokine, for example increased risk of TB and pneumonia.
- Increased levels of antibodies to double-stranded DNA; rare cases of drug-induced lupus can occur.
- Increased risk of lymphomas (not proven).

Differences between TNF-blocking drugs

- Etanercept blocks TNF and lymphotoxin a (LTa).
- Infliximab and adalimumab, but not etanercept, are active in Crohn's disease.
- Difference most likely to be due to different dosing regimes.
- Alleged differences in cytotoxicity/apoptosis are controversial.

Table 1 | Therapeutics for human autoimmunity

Target/therapeutic	Status of therapeutic	Disease outcome	Disadvantages	References
Cytokines				
TNF-specific monoclonal antibody	Approved for rheumatoid arthritis, Crohn's disease, psoriatic arthritis and ankylosing spondylitis	Improvement in disability in all diseases; joint repair in rheumatoid arthritis	Increased risk of TB and other infections; slight increased risk of lymphoma	28–30, 32
Soluble TNFR fusion protein	Approved for rheumatoid arthritis, psoriasis and ankylosing spondylitis	Clinical benefit is the same as TNF-specific monoclonal antibody	Risks are the same as TNF-specific monoclonal antibody therapy	31, 32
IL-1-receptor antagonist	Approved for rheumatoid arthritis	Improves disability	Relatively low efficacy Daily injection	34, 35
IL-15-specific monoclonal antibody	Phase II trial for rheumatoid arthritis	Promising results for disability	Potential for opportunistic infection (blocks natural killer (NK) cells, CD8 memory)	36
IL-6-receptor-specific monoclonal antibody	Phase II trial for rheumatoid arthritis	Decreased disease activity	Potential for opportunistic infection	71
Recombinant type 1 interferons	Approved for relapsing/remitting multiple sclerosis	Reduction in relapse rate	Liver toxicity; influenza-virus like syndrome is common	59
Integrins				
$\alpha_4\beta_1$ -integrin-specific monoclonal antibody	Approved for relapsing/remitting multiple sclerosis Phase II/III trials for rheumatoid arthritis and inflammatory bowel disease	Reduction in relapse rate; delay in progression of disability at two years; encephalopathy	Increased risk of infection Progressive multifocal encephalopathy	6, 10
Oral small-molecule inhibitors	Phase I trials in progress	Not yet known		
HMG-coenzyme A reductase				
Statins	Phase II trials for multiple sclerosis	Reduced activity on magnetic resonance scans	Hepatotoxicity, rhabdomyolysis	72
T cells				
CD3-specific monoclonal antibody	Phase II trials for type 1 diabetes	Reduced insulin usage	Increased risk of infection	14–16
CTLA4-immunoglobulin recombinant protein	Phase III trials for rheumatoid arthritis, psoriasis and multiple sclerosis	Improvement in rheumatoid arthritis		20
B cells				
CD20-specific monoclonal antibody	Phase II trials for rheumatoid arthritis, systemic lupus erythematosus (SLE) and multiple sclerosis	Improvement in rheumatoid arthritis SLE (although extensive co-medication makes interpretation problematic)	Possible increased risk of infection especially if re-treated	24
Antigen-specific T-cell responses				
Random copolymer glatiramer acetate	Approved for relapsing/remitting multiple sclerosis	Reduction in relapse rate	Allergic reactions in 10% of patients	44, 45
Altered peptide ligand to MBP peptide 83–99	Phase IIb trials for multiple sclerosis	Reduced brain lesions (at low doses)	Can exacerbate disease at high doses	46
Altered peptide ligand to HSP60 peptide	Phase II trials for type 1 diabetes	Reduced insulin usage	Allergic reactions in 10% of patients	11
Altered peptide ligand to insulin peptide	Phase II trial in progress for type 1 diabetes	Not yet known	Not yet known	
MBP-encoding tolerizing DNA vaccine	Phase I/II trial in progress for relapsing/remitting multiple sclerosis	Not yet known	Not yet known	11, 48

MBP APL ameliorated EAE in mice induced by a different myelin protein (proteolipid protein), even when the APL was administered after the initial attack of paralysis⁴³. And APL administration similarly induced a shift to T_H2 -cytokine production, reduced epitope spreading, and reduced the broadening of the adaptive T- and B-cell responses. In a phase II placebo-controlled human clinical trial, MBP APL (given in weekly subcutaneous doses) shifted the response of MBP-specific T cells, promoting T_H2 -cytokine production (including IL-4, IL-5, IL-10 and IL-13) and downregulating T_H1 -cytokine production (including IFN- γ and TNF)⁴⁴. Lower doses of MBP APL reduced both the number and the volume of brain lesions detected with magnetic resonance imaging (MRI), but higher doses exacerbated disease in three patients and increased brain lesions⁴⁵. A phase IIb trial is now underway using the lower dose. Three other trials of antigen-specific therapy are underway or recently completed for type 1 diabetes mellitus (T1DM), including phase II trials with glutamic acid decarboxylase, and trials with APLs of

an insulin peptide or of a heat-shock protein 60 (HSP60) peptide. In the trial with the APL of HSP60, decreased exogenous insulin use was observed in diabetics, as well as a T_H2 shift^{46,47}.

An alternative method of targeting antigen-specific responses has recently been developed using DNA constructs that are designed to promote the tolerization of immune responses to multiple myelin components. These DNA constructs encode several myelin antigens, where immune stimulatory motifs (CpG motifs) in the DNA, which promote expression of costimulatory molecules (such as CD28), are replaced with immunosuppressive motifs (GpG motifs), leading to sub-optimal costimulation of antigen-specific T cells (Fig. 2; ref. 48). When administered to mice after the first signs of EAE, these DNA plasmids reduced the subsequent relapse rate over the next three months by more than 50%, and also reduced the spreading of autoantibody responses. A phase I trial with DNA vaccines designed to tolerate against myelin proteins is currently underway.

Oral administration of myelin antigens in multiple sclerosis, collagen in rheumatoid arthritis and insulin in T1DM (which has been shown to favour tolerization of immune responses) has been tested. Despite successfully preventing disease in animal models when antigen was fed at the time of disease induction^{49,50}, clinical trials attempting to treat ongoing disease have been unsuccessful⁵¹. A summary of therapeutics is in Table 1.

Current tools for immunotherapy

Monoclonal antibodies

The success of monoclonal antibodies was slow to arrive, but in 2004, there were two 'blockbusters' on the market (each generating over \$1 billion) — infliximab (Remicade), an anti-TNF antibody, and rituximab, an anti-CD20 antibody. More are on the way; currently almost half of all drug candidates in clinical development are monoclonal antibodies. Infliximab and rituximab are derived from early monoclonal antibody technology. They are 'chimaeric' antibodies, consisting of a mouse combining site (Fv) while the rest (about 70%) is human⁵². Subsequent developments have led to 'humanized' antibodies, in which mouse-derived variable regions (or complementarity-determining regions, CDRs) are grafted into a human antibody scaffold, and 'fully human' antibodies, which contain human variable-region components selected by phage display⁵³. Humanized monoclonal antibodies in the clinic include natalizumab, which blocks $\alpha_4\beta_1$ (ref. 6), and the fully human anti-TNF antibody adalimumab³⁰ (Humira).

Because many potential therapeutic targets are exposed in extracellular fluids (cytokines, chemokines, receptors, other cell-surface molecules and adhesion molecules), they are readily accessible to high-affinity neutralizing antibodies. Furthermore, as natural-body constituents (in contrast to the small-molecule chemicals commonly used as pharmaceuticals), antibodies intrinsically lack toxicity when manufactured, purified and handled properly. Therefore, any toxicity that does occur with monoclonal antibodies is likely to be mechanism related. Another benefit of monoclonal antibodies lies in the fact that even partially humanized antibodies (such as chimaeric antibodies of mouse Fv on a human backbone), as well as fully humanized antibodies, are relatively non-immunogenic. This is probably due to the phenomenon of 'high zone tolerance' described in the 1960s and 1970s that occurs with deaggregated human immunoglobulins⁵⁴ (whereby intravenous deaggregated gammaglobulin was tolerogenic if given in high doses) and the concomitant use of methotrexate, which has immunosuppressive as well as autoinflammatory effects²⁹.

Receptor fusion proteins

Receptor fusion proteins are proteins in which the binding site of a receptor is fused onto an antibody Fc region, which improves the protein's half-life and other pharmacological properties. The most successful receptor fusion protein is etanercept, a dimeric p75 TNFR-immunoglobulin G (IgG) Fc fusion protein³¹ (with sales of over \$1 billion). The clinical benefit of etanercept is indistinguishable from that of anti-TNF antibodies in rheumatoid arthritis^{32,55}, psoriasis and ankylosing spondylitis, although anti-TNF antibodies are more effective in the treatment of inflammatory bowel disease. Receptor fusion proteins are more expensive to manufacture than antibodies, and the use of natural receptors provides for less diversity than with antibodies.

Cytokines

Cytokines have some useful 'drug-like' properties, such as potency, but also some disadvantages, such as a short half-life. But the main problem with cytokines is that they have multiple effects on many cell types²⁵, so systemic injection of cytokines can cause undesirable effects. Thus, the efficacy in animal models of the endogenous anti-inflammatory cytokines IL-10 (ref. 56), IL-4, IL-11 and TGF- β has not translated into their use as human therapeutics, owing to their toxicity. However, the local regulated delivery of cytokines using gene therapy could make them effective as treatments. Recently, it has become possible to engineer cytokines that have enhanced half-lives

Box 3 | Combination drug therapy in serious diseases

Because we do not know the cause of chronic autoimmune diseases, it is unlikely that any single therapy can halt or reverse all the troubling manifestations of these diseases. The way that candidate therapies are often tested — in isolation — predisposes such therapies to failure: in isolation, their effect on a highly complex multifactorial disease process may be relatively small.

Clinically, there has been marked success in the treatment of rheumatoid arthritis by combining methotrexate — an anti-proliferative folic acid inhibitor that inhibits T cells (and other cells) — with TNF inhibitory drugs²⁹. This has been followed by combining methotrexate with other therapeutics, including anti-IL-6R (ref. 33) antibody and CTLA4-immunoglobulin fusion protein²⁰. Methotrexate in combination with anti-TNF therapy was used in an attempt to mimic the augmented benefit of anti-CD4 and anti-TNF antibodies⁷⁹. The lesson here, as in cancer therapeutics, is that more clinical efficacy (and less toxicity) may result from partially blocking several pathways than from complete blockade of any one pathway, which in humans is unattainable. However, certain combinations may be risky. For example, blocking TNF and IL-1 augments the risk of infection⁸⁰ and so caution is necessary to avoid diminishing the benefit:risk ratio.

It is likely that as we understand more about the rate-limiting steps or the 'tipping points' in disease processes, better combinations will be devised to maximise efficacy and to minimize side-effects, the duration of treatment and its cost.

and are activated only at a desired location^{57,58}. Such modifications may overcome some of the inherent difficulties of cytokine therapy.

The type 1 interferons, IFN- α and IFN- β , are effective drugs and have been approved for use in viral infections, some cancers and multiple sclerosis. In multiple sclerosis, relapse rates are reduced by 30% with the administration of IFN- β (ref. 59). However, flu-like symptoms are common during therapy with IFNs, and the immunogenicity of IFNs (probably mechanism related because they upregulate antigen presentation) can limit their efficacy. IFN- β inhibits the activity of metalloproteases 2 and 9. This protease activity is required for lymphocyte homing, so when the administration of IFN- β is combined with adhesion cell blockade, lymphocyte entry into an organ may be drastically reduced¹¹. In this circumstance, endogenous viruses like JC virus, which causes progressive multifocal leukoencephalopathy, may become activated with fatal consequences.

Mutated versions of cytokines can be used as decoys, inhibiting the ability of the endogenous cytokine to act on its receptor. This has been reported with TNF variants that bind to non-mutated endogenous TNF, with the resulting trimeric complex unable to activate TNFRs. In animal models, these TNF variants are effective⁶⁰.

Overcoming limitations

Although there is a lot of optimism among some circles that many new safe therapies are just around the corner, this hope belies the fact that clinical successes, where the benefits outweigh the risks, are few and far between. The failures include antibodies specific for cell-surface antigens such as CD4 and CD25, cytokines such as IL-8, fusion proteins such as the IL-1-receptor 'trap' and the TNFRp55-immunoglobulin fusion protein lenercept, and multiple antigen-specific approaches.

It is thus comforting that there are some clear successes, such as TNF blockade, that are now well established (Table 1). However, the recent withdrawal of anti- $\alpha_4\beta_1$ integrin emphasizes the complexity of reversing ongoing autoimmune disease, without provoking serious complications. Understanding the risk versus benefit relation requires more time than is usually spent in pre-clinical models, and often takes thousands of patient-years of experience to be established.

As summarized in Box 2, anti-TNF therapy of rheumatoid arthritis has marked clinical benefit, with some changes, such as reduction in tiredness, occurring within hours. This benefit occurs in most patients whose condition has not improved following other treatments, such as methotrexate. However, the degree of clinical benefit can vary considerably from patient to patient. The greatest benefit is

seen with combination therapy (Box 3). On the basis of single parameters only, such as joint swelling, all patients improve to some degree²⁸, but if compound parameters are monitored, such as the American College of Rheumatology (ACR) criteria (including number of swollen and tender joints, and levels of C-reactive protein), response rates vary between 50–60% in late-stage disease²⁹, to more than 80% in the early-stage disease⁶¹. In the early stage of disease, there is evidence of disease remissions, which may persist for a year or more after the withdrawal of anti-TNF therapy⁶¹ (F. Breedveld *et al.*, unpublished observations). So early treatment may be the most beneficial and cost-effective. But there is, as yet, no evidence of a cure.

Anti-TNF therapy reduces joint pathology, even in patients showing no clear benefit according to ACR criteria. This suggests that the links between inflammation and joint damage are not fully understood. Most importantly, recent studies have documented evidence for joint repair, after TNF blockade: joint X-rays taken after one year of treatment show an improvement in joint condition compared with those taken before treatment^{62,63}. It is the first example of therapy promoting endogenous repair in any reported human disease.

The most predictable problem of therapy with TNF blockade (and most other immunotherapies including anti- $\alpha_4\beta_1$ -integrin antibody) is augmentation of the risk of infection. In this case, the magnitude of this risk is hard to measure because rheumatoid arthritis patients are more susceptible to infections, partly owing to the disease and partly because of other treatments. The initial incidence of tuberculosis (TB)⁶⁴ in one in every 2,000 patients treated with anti-TNF has been reduced markedly by screening and, if necessary, administration of prophylactic therapy. Other opportunistic infections are rarer, but like TB can occasionally be lethal. More common are respiratory infections. The consensus at present is that the benefit of using TNF blockade in autoimmune diseases with a bad prognosis outweighs the risks^{65,66}.

The risk of infection could be reduced if the duration of TNF blockade were briefer; for example, by using small-molecule chemicals of short half-life. The dilemma here, however, is to define the right therapeutic target. Attempts so far to develop inhibitors of p38 MAP kinase — a component of pro-inflammatory signalling cascades and a favourite target among pharmaceutical companies — have not succeeded, owing to toxicity. Other interesting small-molecule targets, such as IKK2 (inhibitor of NF- κ B kinase 2), are also risky choices because of their presence in almost all cells. Another approach is to target the mechanism involved in the production of TNF in the joints versus that involved in the production of TNF in the immune system, but despite evidence that the mechanism differs, we do not know the molecular targets⁶⁷.

Another common side effect of TNF blockade is the induction of IgM anti-nuclear antibodies, which have been detected in many patients (15%; ref. 68), although IgG antibodies and drug-induced lupus (an antibody-mediated disease) only rarely occur (less than one in 1,000 patients). If lupus does occur, it is reversible, treatable and not nephrotoxic, so is not a major clinical problem.

Lymphomas are more frequent in patients with rheumatoid arthritis than in the normal population, especially those with severe long-standing disease. However, as severe disease is treated by anti-TNF therapy, it is not yet clear whether there is an increased risk of developing lymphomas⁶⁹ after anti-TNF therapy.

Benefit from anti-TNF blockade is not seen in all autoimmune diseases. In fact, the treatment of multiple sclerosis patients with TNF blockade, using lenercept, a TNFRp55-Fc construct which never reached the market, exacerbates the frequency of disease relapse⁸, possibly by augmenting T-cell activity⁷⁰. This discordance may be explained, in part, by the inability of the TNFR fusion constructs to penetrate the inflamed brain owing to the endothelial blood–brain barrier. Alternatively, although TNF may have a destructive role in inflammation in the brain, it may also act as a growth factor for myelin-producing cells, indicating that TNF, similar to many other cytokines, has both harmful and beneficial effects¹¹.

Outlook

Two decades of work defining the molecular basis of the immune response is starting to pay off in the field of autoimmunity. A whole set of 'targeted therapies' has been developed to block many steps in the immune and pro-inflammatory response. Of these, several successes have had a profound impact on patients, on our understanding of disease mechanisms and even on the pharmaceutical industry. The variety of potential therapeutic targets is enormous, but we do not know the rules that define targets of various quality, in terms of their efficacy as well as safety. Some molecules and pathways are common in many autoimmune diseases. TNF is the best-documented example. Hence, TNF blockade is an approved therapy for multiple chronic inflammatory diseases — rheumatoid arthritis, Crohn's disease, psoriatic arthritis, ankylosing spondylitis and psoriasis, with more likely to follow.

The future goal will be to improve the efficacy of immunotherapy, from the current state of partial disease control, to increased disease control, to establishing remission and eventually to cure, without increasing either the risks or costs of treatment. An important step in this progression will be achieving earlier treatment.

For now, the non-antigen-specific approaches are the ones yielding clinical benefit, with the blocking of cytokines, and possibly adhesion molecules, being the most effective. But with such non-antigen-specific approaches, the risk of opportunistic infection is problematic. In the future, non-antigen-specific approaches may be made safer by targeting them to the site of disease, for example by gene therapy. But the most obvious way to reduce opportunistic infections is to use antigen-specific therapy — a dream of immunologists for generations now. Although several attempts in the past decade have failed, we are optimistic that eventually, the molecular understanding of tolerance and immunity will progress, and the 'holy grail' of autoimmunity — long-term antigen-specific therapy — will be reached. The progress made in devising rational and effective non-antigen-specific therapy reflects the development of useful research and therapeutic tools, and provides grounds for this optimism. ■

1. Sayegh, M. H. & Carpenter, C. B. Transplantation 50 years later — progress, challenges, and promises. *N. Engl. J. Med.* **351**, 2761–2766 (2004).
2. Roep, B. O., Atkinson, M. & von Herrath, M. Satisfaction (not) guaranteed: re-evaluating the use of animal models of type 1 diabetes. *Nature Rev. Immunol.* **4**, 989–997 (2004).
3. Bach, J. F. Immunotherapy of type 1 diabetes: lessons for other autoimmune diseases. *Arthritis Res.* **4** (suppl. 3), S3–S15 (2002).
4. Malfait, A. M., Williams, R. O., Malik, A. S., Maini, R. N. & Feldmann, M. Chronic relapsing homologous collagen-induced arthritis in DBA/1 mice as a model for testing disease-modifying and remission-inducing therapies. *Arthritis Rheum.* **44**, 1215–1224 (2001).
5. Feldmann, M. Development of anti-TNF therapy for rheumatoid arthritis. *Nature Rev. Immunol.* **2**, 364–371 (2002).
6. Miller, D. H. *et al.* A controlled trial of natalizumab for relapsing multiple sclerosis. *N. Engl. J. Med.* **348**, 15–23 (2003).
7. Gladwell, M. *The Tipping Point: How Little Things Can Make a Difference* (Little, Brown & Co., Boston, 2000).
8. Feldmann, M. & Maini, R. N. Lasker clinical medical research award. TNF defined as a therapeutic target for rheumatoid arthritis and other autoimmune diseases. *Nature Med.* **9**, 1245–1250 (2003).
9. Yednock, T. A. *et al.* Prevention of experimental autoimmune encephalomyelitis by antibodies against $\alpha_4\beta_1$ integrin. *Nature* **356**, 63–66 (1992).
10. Ghosh, S. *et al.* Natalizumab for active Crohn's disease. *N. Engl. J. Med.* **348**, 24–32 (2003).
11. Steinman, L. Immune therapy for autoimmune diseases. *Science* **305**, 212–216 (2004).
12. Breedveld, F. C. Monoclonal antibodies to CD4. *Rheum. Dis. Clin. North Am.* **24**, 567–578 (1998).
13. Lindsey, J. W. *et al.* Repeated treatment with chimeric anti-CD4 antibody in multiple sclerosis. *Ann. Neurol.* **36**, 183–189 (1994).
14. Chatenoud, L. CD3-specific antibody-induced active tolerance: from bench to bedside. *Nature Rev. Immunol.* **3**, 123–132 (2003).
15. Herold, K. C. *et al.* Anti-CD3 monoclonal antibody in new-onset type 1 diabetes mellitus. *N. Engl. J. Med.* **346**, 1692–1698 (2002).
16. Charpentier, B. *et al.* Evidence that antihuman tumor necrosis factor monoclonal antibody prevents OKT3-induced acute syndrome. *Transplantation* **54**, 997–1002 (1992).
17. Banchereau, J. & Steinman, R. M. Dendritic cells and the control of immunity. *Nature* **392**, 245–252 (1998).
18. Bottazzo, G. F., Pujol-Borrell, R., Hanafusa, T. & Feldmann, M. Role of aberrant HLA-DR expression and antigen presentation in induction of endocrine autoimmunity. *Lancet* **2**, 1115–1119 (1983).
19. McDevitt, H. O., Perry, R. & Steinman, L. A. Monoclonal anti-Ia antibody therapy in animal models of autoimmune disease. *Ciba Found. Symp.* **129**, 184–193 (1987).
20. Kremer, J. M. *et al.* Treatment of rheumatoid arthritis by selective inhibition of T-cell activation with fusion protein CTLA4Ig. *N. Engl. J. Med.* **349**, 1907–1915 (2003).
21. Humphreys, I. R. *et al.* A critical role for OX40 in T cell-mediated immunopathology during lung viral infection. *J. Exp. Med.* **198**, 1237–1242 (2003).

22. Shevach, E. M. Regulatory/suppressor T cells in health and disease. *Arthritis Rheum.* **50**, 2721–2724 (2004).
23. Bluestone, J. A. & Tang, Q. Therapeutic vaccination using CD4⁺CD25⁺ antigen-specific regulatory T cells. *Proc. Natl. Acad. Sci. USA* **101** (suppl. 2), 14622–14626 (2004).
24. Kazkaz, H. & Isenberg, D. Anti B cell therapy (rituximab) in the treatment of autoimmune diseases. *Curr. Opin. Pharmacol.* **4**, 398–402 (2004).
25. Oppenheim, J. J. & Feldmann, M. in *Cytokine Reference, Vol. 1: Ligands* (eds Oppenheim, J. J. & Feldmann, M.) 3–20 (Academic, London, 2001).
26. Feldmann, M. & Brennan, F. M. in *Cytokine Reference, Vol. 1: Ligands* (eds Oppenheim, J. J. & Feldmann, M.) 35–41 (Academic, London, 2001).
27. Feldmann, M., Brennan, F. M. & Maini, R. N. Role of cytokines in rheumatoid arthritis. *Annu. Rev. Immunol.* **14**, 397–440 (1996).
28. Elliott, M. J. et al. Treatment of rheumatoid arthritis with chimeric monoclonal antibodies to tumor necrosis factor alpha. *Arthritis Rheum.* **36**, 1681–1690 (1993).
29. Maini, R. N. et al. Therapeutic efficacy of multiple intravenous infusions of anti-tumor necrosis factor alpha monoclonal antibody combined with low-dose weekly methotrexate in rheumatoid arthritis. *Arthritis Rheum.* **41**, 1552–1563 (1998).
30. Keystone, E. C. et al. Radiographic, clinical, and functional outcomes of treatment with adalimumab (a human anti-tumor necrosis factor monoclonal antibody) in patients with active rheumatoid arthritis receiving concomitant methotrexate therapy: a randomized, placebo-controlled, 52-week trial. *Arthritis Rheum.* **50**, 1400–1411 (2004).
31. Moreland, L. W. et al. Treatment of rheumatoid arthritis with a recombinant human tumor necrosis factor receptor (p75)-Fc fusion protein. *N. Engl. J. Med.* **337**, 141–147 (1997).
32. Feldmann, M. & Maini, R. N. Anti-TNFα therapy of rheumatoid arthritis: what have we learned? *Annu. Rev. Immunol.* **19**, 163–196 (2001).
33. Nishimoto, N. & Kishimoto, T. Inhibition of IL-6 for the treatment of inflammatory diseases. *Curr. Opin. Pharmacol.* **4**, 386–391 (2004).
34. Butler, D. M., Maini, R. N., Feldmann, M. & Brennan, F. M. Modulation of proinflammatory cytokine release in rheumatoid synovial membrane cell cultures. Comparison of monoclonal anti TNF-α antibody with the interleukin-1 receptor antagonist. *Eur. Cytokine Netw.* **6**, 225–230 (1995).
35. Bresnahan, B. et al. Treatment of rheumatoid arthritis with recombinant human interleukin-1 receptor antagonist. *Arthritis Rheum.* **41**, 2196–2204 (1998).
36. McInnes, I. B. & Gracie, J. A. Interleukin-15: a new cytokine target for the treatment of inflammatory diseases. *Curr. Opin. Pharmacol.* **4**, 392–397 (2004).
37. Czura, C. J., Yang, H., Amella, C. A. & Tracey, K. J. HMGB1 in the immunology of sepsis (not septic shock) and arthritis. *Adv. Immunol.* **84**, 181–200 (2004).
38. Bekker, P. J. et al. A single-dose placebo-controlled study of AMG162, a fully human monoclonal antibody to RANKL, in postmenopausal women. *J. Bone Miner. Res.* **19**, 1059–1066 (2004).
39. Baggiolini, M. Reflections on chemokines. *Immunol. Rev.* **177**, 5–7 (2000).
40. Springer, T. A. Traffic signals for lymphocyte recirculation and leukocyte emigration: the multistep paradigm. *Cell* **76**, 301–314 (1994).
41. Yu, M., Johnson, J. M. & Tuohy, V. K. A predictable sequential determinant spreading cascade invariably accompanies progression of experimental autoimmune encephalomyelitis: a basis for peptide-specific therapy after onset of clinical disease. *J. Exp. Med.* **183**, 1777–1788 (1996).
42. Sela, M. The concept of specific immune treatment against autoimmune diseases. *Int. Rev. Immunol.* **18**, 201–216 (1999).
43. Brocke, S. et al. Treatment of experimental encephalomyelitis with a peptide analogue of myelin basic protein. *Nature* **379**, 343–346 (1996).
44. Kappos, L. et al. Induction of a non-encephalitogenic type 2 T helper-cell autoimmune response in multiple sclerosis after administration of an altered peptide ligand in a placebo-controlled, randomized phase II trial. The altered peptide ligand in relapsing MS study group. *Nature Med.* **6**, 1176–1182 (2000).
45. Bielekova, B. et al. Encephalitogenic potential of the myelin basic protein peptide (amino acids 83–99) in multiple sclerosis: results of a phase II clinical trial with an altered peptide ligand. *Nature Med.* **6**, 1167–1175 (2000).
46. Ruiz, P. J. et al. Suppressive immunization with DNA encoding a self-peptide prevents autoimmune disease: modulation of T cell co-stimulation. *J. Immunol.* **162**, 3336–3341 (1999).
47. Raz, I. et al. Beta-cell function in new-onset type 1 diabetes and immunomodulation with a heat-shock protein peptide (DiaPep277): a randomised, double-blind, phase II trial. *Lancet* **358**, 1749–1753 (2001).
48. Robinson, W. H. et al. Protein microarrays guide tolerizing DNA vaccine treatment of autoimmune encephalomyelitis. *Nature Biotechnol.* **21**, 1033–1039 (2003).
49. Gutgemann, I., Fahrner, A. M., Altman, J. D., Davis, M. M. & Chien, Y. H. Induction of rapid T cell activation and tolerance by systemic presentation of an orally administered antigen. *Immunity* **8**, 667–673 (1998).
50. Chen, Y., Kuchroo, V. K., Inobe, J., Hafler, D. A. & Weiner, H. L. Regulatory T cell clones induced by oral tolerance: suppression of autoimmune encephalomyelitis. *Science* **265**, 1237–1240 (1994).
51. Toussiot, E. A. Oral tolerance in the treatment of rheumatoid arthritis. *Curr. Drug Targets Inflamm. Allergy* **1**, 45–52 (2002).
52. Knight, D. M. et al. Construction and initial characterization of a mouse-human chimeric anti-TNF antibody. *Mol. Immunol.* **30**, 1443–1453 (1993).
53. Winter, G., Griffiths, A. D., Hawkins, R. E. & Hoogenboom, H. R. Making antibodies by phage display technology. *Annu. Rev. Immunol.* **12**, 433–455 (1994).
54. Chiller, J. M., Habicht, G. S. & Weigle, W. O. Cellular sites of immunologic unresponsiveness. *Proc. Natl. Acad. Sci. USA* **65**, 551–556 (1970).
55. Bathon, P. M. et al. A comparison of etanercept and methotrexate in patients with early rheumatoid arthritis. *N. Engl. J. Med.* **343**, 1586–1593 (2000).
56. Walmsley, M. et al. Interleukin-10 inhibition of the progression of established collagen-induced arthritis. *Arthritis Rheum.* **39**, 495–503 (1996).
57. Adams, G., Vessillier, S., Dreja, H. & Chernajovsky, Y. Targeting cytokines to inflammation sites. *Nature Biotechnol.* **21**, 1314–1320 (2003).
58. Steinman, L. Engineering better cytokines. *Nature Biotechnol.* **21**, 1293–1294 (2003).
59. Revel, M. Interferon-beta in the treatment of relapsing-remitting multiple sclerosis. *Pharmacol. Ther.* **100**, 49–62 (2003).
60. Steed, P. M. et al. Inactivation of TNF signaling by rationally designed dominant-negative TNF variants. *Science* **301**, 1895–1898 (2003).
61. Quinn, M. A. et al. Very early treatment with infliximab in addition to methotrexate in early, poor-prognosis rheumatoid arthritis reduces magnetic resonance imaging evidence of synovitis and damage, with sustained benefit after infliximab withdrawal: results from a twelve-month randomized, double-blind, placebo-controlled trial. *Arthritis Rheum.* **52**, 27–35 (2005).
62. Lipsky, P. E. et al. Infliximab and methotrexate in the treatment of rheumatoid arthritis. Anti-tumor necrosis factor trial in rheumatoid arthritis with concomitant therapy study group. *N. Engl. J. Med.* **343**, 1594–1602 (2000).
63. Klareskog, L. et al. Therapeutic effect of the combination of etanercept and methotrexate compared with each treatment alone in patients with rheumatoid arthritis: double-blind randomised controlled trial. *Lancet* **363**, 675–681 (2004).
64. Keane, J. et al. Tuberculosis associated with infliximab, a tumor necrosis factor α-neutralizing agent. *N. Engl. J. Med.* **345**, 1098–1104 (2001).
65. Day, R. Adverse reactions to TNF-α inhibitors in rheumatoid arthritis. *Lancet* **359**, 540–541 (2002).
66. Pisetsky, D. S. & St Clair, E. W. Progress in the treatment of rheumatoid arthritis. *JAMA* **286**, 2787–2790 (2001).
67. Brennan, F. M. et al. Evidence that rheumatoid arthritis synovial T cells are similar to cytokine-activated T cells. *Arthritis Rheum.* **46**, 31–41 (2002).
68. Charles, P. J., Smeenk, R. J. T., DeJong, J., Feldmann, M. & Maini, R. N. Assessment of antibodies to double-stranded DNA induced in rheumatoid arthritis patients following treatment with infliximab, a monoclonal antibody to tumor necrosis factor α. *Arthritis Rheum.* **43**, 2383–2390 (2000).
69. Baecklund, E., Askling, J., Rosenquist, R., Ekbo, A. & Klareskog, L. Rheumatoid arthritis and malignant lymphomas. *Curr. Opin. Rheumatol.* **16**, 254–261 (2004).
70. Cope, A. P. et al. Chronic exposure to tumor necrosis factor (TNF) *in vitro* impairs the activation of T cells through the T cell receptor/CD3 complex; reversal *in vivo* by anti-TNF antibodies in patients with rheumatoid arthritis. *J. Clin. Invest.* **94**, 749–760 (1994).
71. Nishimoto, N. et al. Treatment of rheumatoid arthritis with humanized anti-interleukin-6 receptor antibody: a multicenter, double-blind, placebo-controlled trial. *Arthritis Rheum.* **50**, 1761–1769 (2004).
72. Vollmer, T. et al. Oral simvastatin treatment in relapsing-remitting multiple sclerosis. *Lancet* **363**, 1607–1608 (2004).
73. Kwak, B., Mulhaupt, F., Myit, S. & Mach, F. Statins as a newly recognized type of immunomodulator. *Nature Med.* **6**, 1399–1402 (2000).
74. Garren, H. et al. Combination of gene delivery and DNA vaccination to protect from and reverse Th1 autoimmune disease via deviation to the Th2 pathway. *Immunity* **15**, 15–22 (2001).
75. Youssef, S. et al. The HMG-CoA reductase inhibitor, atorvastatin, promotes a Th2 bias and reverses paralysis in central nervous system autoimmune disease. *Nature* **420**, 78–84 (2002).
76. McCarey, D. W. et al. Trial of Atorvastatin in Rheumatoid Arthritis (TARA): double-blind, randomised placebo-controlled trial. *Lancet* **363**, 2015–2021 (2004).
77. Lovett-Racke, A. E. et al. Peroxisome proliferator-activated receptor alpha agonists as therapy for autoimmune disease. *J. Immunol.* **172**, 5790–5798 (2004).
78. Dalbeth, N., Edwards, J., Fairchild, S., Callan, M. & Hall, F. C. The non-thiol angiotensin-converting enzyme inhibitor quinapril suppresses inflammatory arthritis. *Rheumatology (Oxford)* **44**, 24–31 (2005).
79. Williams, R. O., Mason, L. J., Feldmann, M. & Maini, R. N. Synergy between anti-CD4 and anti-tumor necrosis factor in the amelioration of established collagen-induced arthritis. *Proc. Natl. Acad. Sci. USA* **91**, 2762–2766 (1994).
80. Genovese, M. C. et al. Combination therapy with etanercept and anakinra in the treatment of patients with rheumatoid arthritis who have been treated unsuccessfully with methotrexate. *Arthritis Rheum.* **50**, 1412–1419 (2004).

Competing interests statement The authors declare competing financial interests: details accompany the paper on www.nature.com/nature

Treatment of severe autoimmune disease by stem-cell transplantation

Megan Sykes¹ and Boris Nikolic²

Transplantation of haematopoietic stem cells — cells capable of self renewing and reconstituting all types of blood cell — can treat numerous lethal diseases, including leukaemias and lymphomas. It may now be applicable for the treatment of severe autoimmune diseases, such as therapy-resistant rheumatoid arthritis and multiple sclerosis. Studies in animal models show that the transfer of haematopoietic stem cells can reverse autoimmunity, and several mechanistic pathways may explain this phenomenon. The outcome of ongoing clinical trials, as well as of studies in patients and animal models, will help to determine the role that stem-cell transplantation can play in the treatment of autoimmune diseases.

Most patients with autoimmune disease have a near-normal life expectancy. Nevertheless, some patients suffer severe, therapy-resistant progressive autoimmunity. Haematopoietic-cell transplantation (HCT) is a potential therapy for people with such severe refractory diseases. HCT involves the administration of haematopoietic stem cells (HSC), which are self-renewing and capable of giving rise to all mature haematopoietic cell types and possibly to some non-haematopoietic cell types. The HSC phenotype is characterized by a lack of cell-surface lineage markers and the expression of CD34 in humans. For the sake of simplicity, in this review we use the general term HCT to encompass the transplantation of mobilized peripheral blood cells and bone-marrow cells, both of which are enriched for HSC, as well as to describe purified HSC. The inclusion of other cells besides HSC, especially lymphocytes, in an HCT product, has considerable potential to influence the outcome of HCT.

The recipient is prepared for the transplant by potent immunosuppressive treatment, usually by chemotherapy and/or radiotherapy. This may then be followed by the transfer of autologous haematopoietic cells (auto-HCT; cells harvested from the recipient before patient conditioning (Fig. 1)) or allogeneic haematopoietic cells (allo-HCT; cells harvested from donors rather than the recipient (Fig. 2)) to restore the host immune system. This procedure can cure autoimmune disease in experimental animal models and is now being explored in human clinical trials. So far, auto-HCT has been generally preferred over allo-HCT because of the increased toxicity and potential for rejection in allo-HCT. The risk of graft-versus-host-disease (GVHD) in allo-HCT, which arises from the attack of donor allogeneic T cells on recipient alloantigens, is associated with significant morbidity and mortality. Nevertheless, allo-HCT has been associated with durable complete remission in a small number of patients treated for malignant diseases with coincidental autoimmune disorders. Most recently, several groups have applied non-myeloablative or reduced-intensity conditioning protocols for allo-HCT to the induction of transplantation tolerance and re-establishment of self-tolerance in animal models of autoimmune disease. Although these recent advances have made allo-HCT less toxic, it is an intensive procedure that most probably will not replace current pharmacological treatment for less severe, conservatively treatable autoimmune diseases.

In this article, we review the mechanisms by which HCT might induce tolerance to allo- and autoantigens and discuss recent advances that provide hope for better treatment of severe refractory autoimmune disease. The potential for tissue regeneration using this process is also discussed.

Animal models provide the rationale for HCT

Susceptibility to autoimmune diseases appears to reside in haematopoietic cells. Transplantation of bone-marrow cells from lupus-prone NZB mice into lethally irradiated non-susceptible strains induced the lupus syndrome¹. This transfer of immune diseases with haematopoietic cells was subsequently confirmed for many autoimmune diseases, including other murine models of systemic lupus erythematosus (SLE), experimental autoimmune encephalomyelitis (EAE), adjuvant arthritis, antiphospholipid syndrome and type 1 diabetes². Resistance to autoimmune diseases could also be transferred to susceptible strains by haematopoietic cells. In mice, HCT from disease-resistant allogeneic animals (allo-HCT) prevented the development of SLE, type 1 diabetes, EAE and other autoimmune diseases^{3,4}.

Although the prevention of autoimmunity might some day be clinically feasible, at the moment we cannot predict such diseases accurately enough to justify the use of toxic preventive treatments. Unfortunately, animal studies show that preventing the onset of autoimmunity is much easier than reversing established disease. As of 2004, for example, more than 195 methods for preventing or delaying the onset of type 1 diabetes in non-obese diabetic (NOD) mice had been identified⁵. But only a few therapeutic approaches have been effective in reversing established SLE and type 1 diabetes in mice, one of which is allo-HCT^{6,7}.

In animals, syngeneic HCT (that is, HCT from a genetically identical donor) can be used instead of auto-HCT. For simplicity we will refer to syngeneic HCT as auto-HCT. Although allo-HCT was reliably effective in reversing autoimmunity in animals, auto-HCT did not successfully reverse type 1 diabetes or SLE. However, varying levels of disease remission were achieved following auto-HCT in models of myasthenia gravis⁸, adjuvant arthritis⁹ and EAE¹⁰. Allo-HCT was superior in the therapy of EAE. Auto-HCT was most successful at early dis-

¹Bone Marrow Transplantation Section, Transplantation Biology Research Center, Massachusetts General Hospital/Harvard Medical School, MGH East, Building 149-5102, 13th Street, Boston, Massachusetts 02129, USA (e-mail: sykes@tbr.mgh.harvard.edu). ²The Renal Unit, Massachusetts General Hospital/Harvard Medical School, MGH East, Building 149-8230, 13th Street, Boston, Massachusetts 02129, USA (e-mail: bnkolic@partners.org).

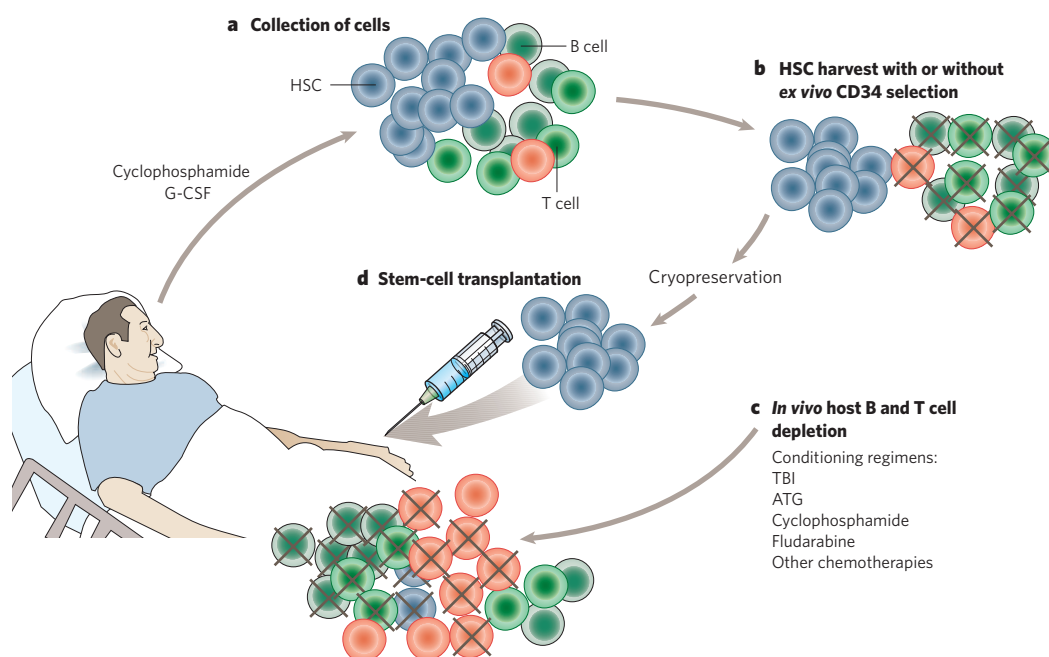


Figure 1 | Autologous HCT as a therapy for autoimmune disease. Red cells represent autoreactive host lymphocytes; green cells represent non-autoreactive host lymphocytes; blue cells represent HSC. The steps in autologous HCT include the following. **a**, Stem-cell collection: HSCs are either harvested from the bone marrow under anaesthesia or collected after mobilization to the peripheral blood by treating the patient with growth factor (for example, G-CSF) and/or chemotherapy (for example, cyclophosphamide). **b**, HSC harvest, with or without *ex vivo* selection using antibody specific for CD34 (a marker of HSCs): efforts may or may not be made to deplete T and B cells from the marrow or

mobilized peripheral blood cells. This is usually achieved by HSC positive selection on the basis of CD34 expression. The HSC preparation is cryopreserved. **c**, *In vivo* lymphocyte depletion: various conditioning regimens with the goal of minimizing the burden of autoreactive B and T cells are given to patients. These include combinations of chemotherapeutic agents, total-body irradiation (TBI) and *in vivo* lymphocyte depletion by antibodies, such as anti-thymocyte globulin (ATG), and these treatments often deplete host HSC as well. **d**, Stem-cell transplantation: after conditioning, the cryopreserved HSC preparation is thawed and returned to the recipient to reconstitute haematopoiesis.

ease stages; nearly complete reversal of disease was achieved, but no effect was observed at more advanced stages^{11,12}. Also, increased intensity of conditioning therapy correlated positively with the success of auto-HCT¹¹.

Several observations in animal models of auto- and allo-HCT suggest that these approaches to the therapy of refractory autoimmune disease will be beneficial for patients. First, the partial clinical success of high doses of chemotherapeutic agents, such as cyclophosphamide, for the treatment of autoimmune diseases hinted that even greater success might be achieved with more intensive immunoablative treatments. The major dose-limiting side effect of these treatments is haematopoietic suppression, and this can be reversed by HCT. Given the complications of GVHD and graft failure following allo-HCT and the positive animal data described above, auto-HCT is a logical approach for haematopoietic rescue. Second, the superior efficacy of allo-HCT in some animal autoimmune disease models suggested that allogeneic donor T cells eliminate autoreactive host lymphocytes, and therefore mediate a form of immunotherapy termed graft-versus-autoimmunity (GVA). There is no reason to expect that such immune responses are specific for autoreactive lymphocytes. Therefore this immune reactivity is likely to be a graft-versus-host (GVH) response. Third, animal studies indicate that resistance to autoimmune disease resides in and can be transferred by HSC²⁻⁴. Finally, engraftment of allo-HSC could result in tolerance to donor antigens, permitting acceptance of donor-derived tissue (for example, islets) without chronic immunosuppressive therapy. Thus, animal studies suggest several potentially useful clinical strategies involving HCT.

Complex, inbred rodent models of autoimmunity provide an opportunity to dissect mechanisms and explore therapies in what is essentially a single donor–recipient combination. But to extend observations from these animal models to humans, several factors must be borne in mind: the impact of the complex and varied genetics of

humans; the effect of different reagents used and different responses to treatment modalities; the impact of concurrent and complicating co-morbidities in the patient; and the ability to regenerate an immune system after lymphoablative therapy. Thus, therapies that are safe and effective in murine models may be neither safe nor effective in human trials, and subjects for such trials require careful selection.

Mechanisms by which HCT might treat autoimmunity

Several mechanisms for correction of autoimmunity may apply to HCT (Fig. 3). These are unlikely to be mutually exclusive.

Immunomodulation by immunosuppressive conditioning and auto-HCT

Potent immunosuppressive treatments can eliminate autoreactive memory B cells and T cells (Fig. 3a), thus ‘resetting’ the immune system. In view of the hypothesized role of microbial antigens and additional unknown environmental triggers in promoting autoimmune disease, it is possible that elimination of the existing T and/or B cell repertoire allows the patient’s immune system to ‘start from scratch’ and regenerate normally. Newly developing B and T cells may be introduced to autoantigens and undergo self-tolerization; alternatively, the chance environmental circumstances that led to the initial autoimmunity might never happen again in the patient’s lifetime. The concordance rate of autoimmune diseases in identical twins can be used to gauge the expected recurrence rates if all pre-existing autoreactive lymphocytes were ablated before auto-HCT. These are in the range 10–35%, depending on the disease, and suggest that this approach could lead to sustained remissions in most patients if complete ablation of autoreactive lymphocytes could be achieved.

However, it is not clear whether autoreactive memory T cells, memory B cells and long-lived plasma cells can be completely removed from the haematolymphoid and target organs by any conditioning treatment. Incomplete immunoablation may account for the subopti-

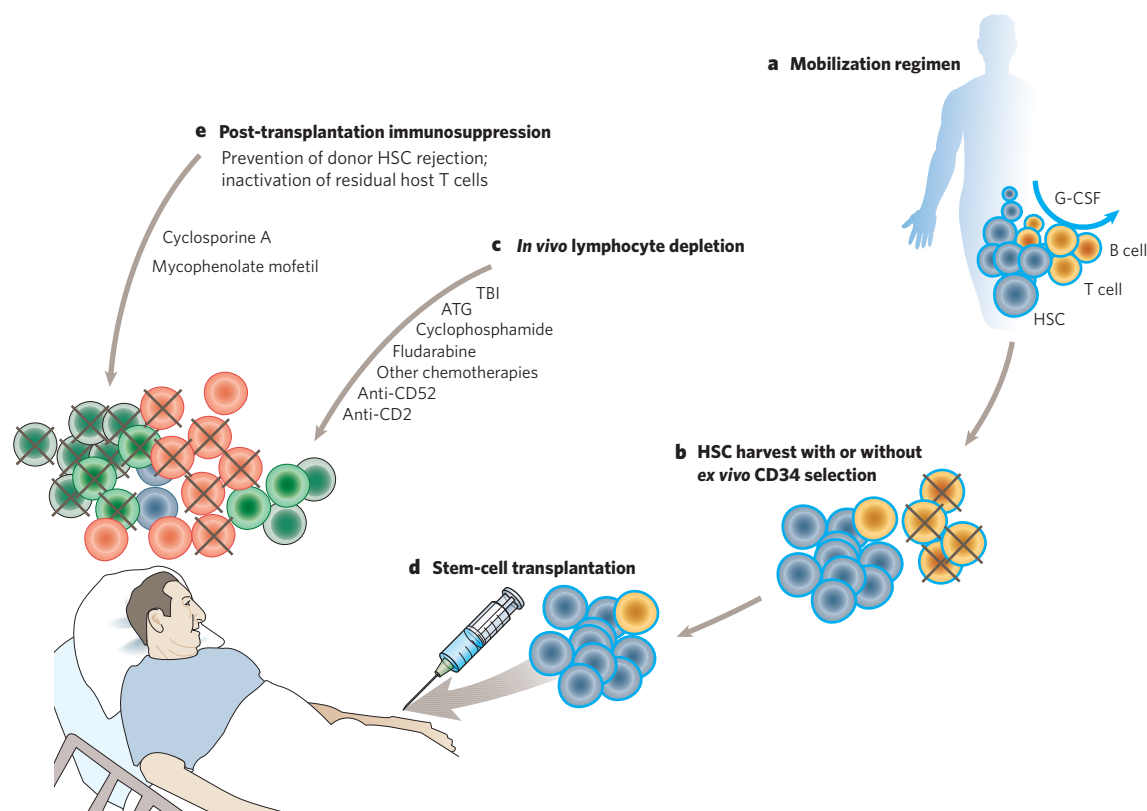


Figure 2 | Allogeneic HCT as a therapy for autoimmune disease. Red cells represent autoreactive host lymphocytes; green cells represent non-autoreactive host lymphocytes; yellow denotes donor lymphocytes; blue cells represent HSC. The steps in allogeneic HCT include the following. **a**, Mobilization regimen: these are similar to the methods employed for autologous HCT, except that cyclophosphamide and other chemotherapies are not used to mobilize stem cells. Mobilization is achieved using the growth factor G-CSF. **b**, HSC harvest, with or without *ex vivo* CD34 selection. This

step is similar to that used for auto-HCT. **c**, *In vivo* lymphocyte depletion: this step depletes host T and B cells to prevent rejection and treat the autoimmune disease; it may also be aimed at depleting donor T cells (to prevent GVHD). **d**, Stem-cell transplantation: as for auto-HCT. **e**, Post-transplantation immunosuppression: as prophylaxis against both rejection mediated by residual host T cells and GVHD mediated by contaminating donor T cells, allo-HCT usually requires additional post-transplantation immunosuppression with agents such as cyclosporine A and mycophenolate mofetil.

mal responses and high rates of early relapse seen in some clinical trials of auto-HCT (see below). Even with incomplete immunoablation, however, the dependence on multiple cell populations for the maintenance of some autoreactions may lead to improvement in disease. For example, intense conditioning chemotherapy induces severe and prolonged depression of CD4⁺ T cells, characterized by an inverted CD4⁺:CD8⁺ T-cell ratio, a predominance of memory T cells and predominant extrathymic pathways of T-cell recovery¹³. These recovering T cells may exhibit increased susceptibility to apoptosis and altered function¹³. Therefore, altered immune reconstitution and non-specific post-conditioning immunosuppression may result in prolonged disease remission.

In animal models and clinical trials, lymphoablative therapies alone (without HCT) can halt or slow the progression of autoimmune diseases. High-dose cyclophosphamide treatment can reduce the number of cells in the mature immune system, resulting in durable remission in severe aplastic anaemia, Felty's syndrome (characterized by rheumatoid arthritis, splenomegaly and granulocytopenia), autoimmune thrombocytopenic purpura and SLE¹⁴. This treatment alone is non-myeloablative and does not require HCT for rescue. Although regimens containing cyclophosphamide are highly effective in SLE, they do not achieve satisfactory responses for rheumatoid arthritis. Further studies are needed to evaluate the optimal conditioning regimen for each autoimmune disease as well as the duration and stability of the induced remissions.

To obtain cells for auto-HCT before conditioning, stem cells are mobilized from the marrow to the peripheral blood, which permits

stem-cell collection without general anaesthesia or bone-marrow harvest. This is achieved by various protocols, many of which involve granulocyte colony-stimulating factor (G-CSF) or cyclophosphamide, a drug that is myelosuppressive but leads to a 'rebound' mobilization of stem cells. Auto-HCT using mobilized peripheral blood stem cells allows strongly immunoablative treatments to be used in patients with refractory autoimmune diseases. Multiple clinical trials of high-dose immunosuppression and chemotherapy with auto-HCT have been initiated in patients with multiple sclerosis, systemic sclerosis, rheumatoid arthritis, SLE and myasthenia gravis^{15–18}. These have involved conditioning regimens of varying intensity, variable *in vivo* and *ex vivo* T-cell purging, and variable stem-cell sources and mobilization protocols. Data from phase I/II clinical trials suggest that auto-HCT is feasible and that it may lead to remission of these diseases^{15–18}. Although the results vary from disease to disease, over a third of patients have sustained a durable remission, often with no further need for immunosuppressive drugs¹⁸.

Much has been learned about the patient subgroups most likely to benefit from this approach, and we refer the reader to several excellent recent reviews on this topic^{16–18}. In some diseases, however, recurrences are relatively frequent^{16–18}. Since controversy prevails about the relative importance of T and B cells in the pathogenesis of some autoimmune diseases (such as rheumatoid arthritis), the eradication of memory B and T cells and tolerization of newly developing cells of each lineage may be more relevant to some diseases than others. Although it is clear that toxicity is greatest with high-intensity conditioning regimens and that infectious risks are increased by T-cell purging owing to delayed

immune reconstitution¹⁸, it is not yet clear whether increased conditioning intensity or lymphocyte purging of the HCT (to avoid reinfusion of autoreactive cells) is associated with improved disease outcomes. Clarification of these issues will await randomized phase III trials that compare auto-HCT with standard therapy. These have recently begun in the United States and Europe. Ideally, these data will help to clarify whether the major benefit of such treatments relates to their ability to deplete autoreactive lymphocytes or to re-establish immune regulation that limits autoimmunity. Furthermore, these results will provide proof of efficacy, which is required for the wider application of auto-HCT in various autoimmune diseases.

Immune-mediated destruction of autoreactive T and B cells

Auto-HCT conditioning regimens for the treatment of leukaemia and lymphoma are designed to destroy the maximum number of malignant cells; the elimination of every single malignant cell is not possible. Total elimination and a cure are often only possible using allo-HCT, where

total elimination results from a GVH immune response, owing to donor T cells recognizing the alloantigens of the recipient. This is often associated with GVHD — a disease of the skin, intestines and liver. Despite its associated morbidity and mortality, GVHD is linked to enhanced graft-versus-leukaemia (GVL) effects, which reflect immune destruction of residual leukaemic cells, and enhanced engraftment, probably owing to destruction of residual host T cells. A similar beneficial effect has been observed using allo-HCT to treat autoimmunity¹⁹. Although auto-HCT permits intense host immune suppression, elimination of every resting memory lymphocyte with high-dose chemotherapy and/or radiation is probably not feasible. Complete immune ablation and reliable reversal of autoimmunity may require the GVA effects of allo-HCT (Fig. 3b), as suggested by the superior results of allo-HCT in numerous animal models. However, this benefit comes with the associated risk of GVHD.

The association of GVA with GVHD was confirmed using a meta-analysis of individual patient data¹⁹ and was seen with discontinuation

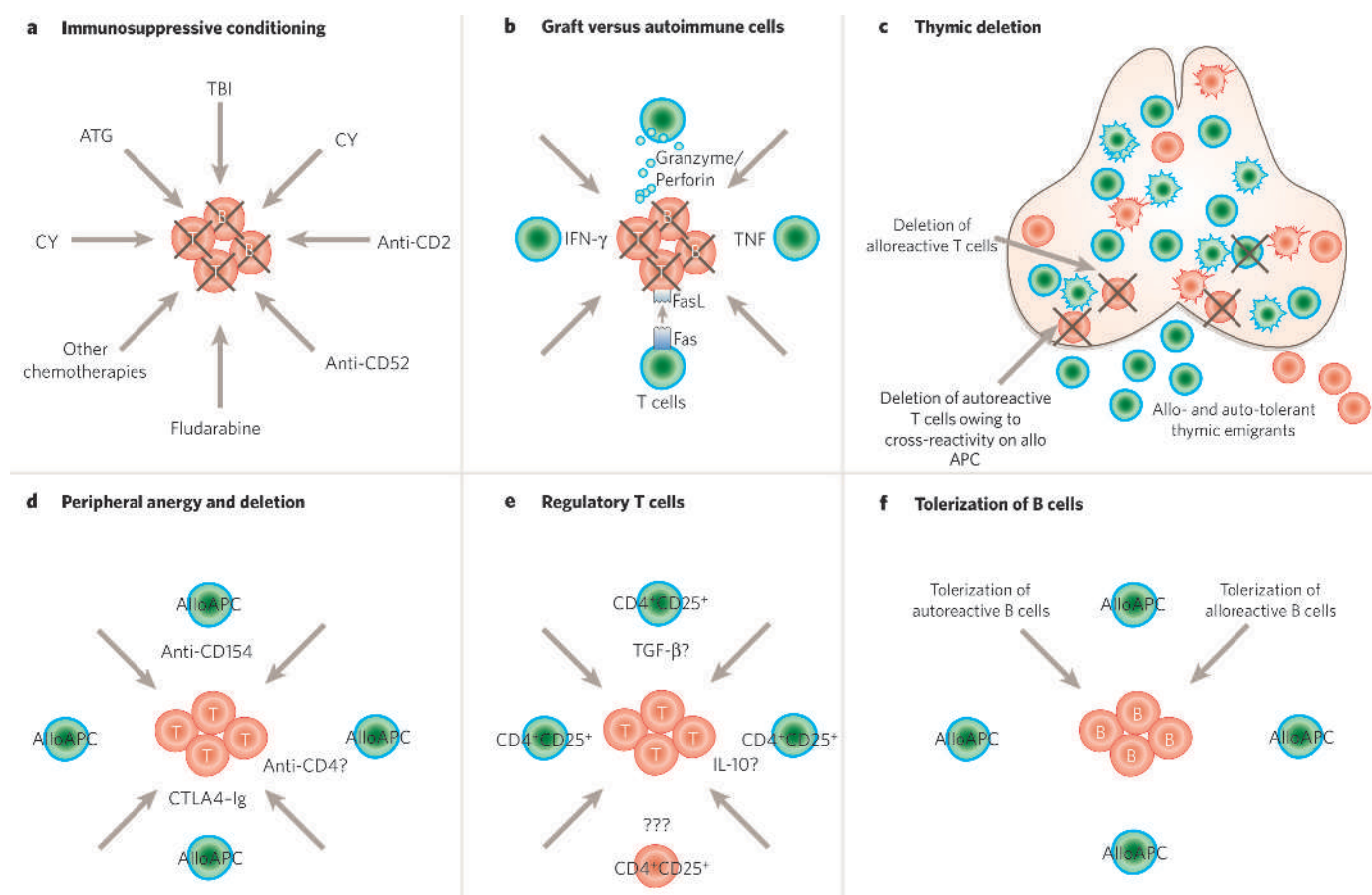


Figure 3 | Mechanisms by which HCT might ameliorate autoimmune diseases. These mechanisms are not mutually exclusive. Donor-derived cells are shown in green and host-derived cells are shown in red.

a, Immunomodulation by immunosuppressive conditioning: potent immunosuppressive treatments such as total body irradiation (TBI), cyclophosphamide (CY), anti-CD2 antibodies (anti-CD2), anti-CD52 antibodies (anti-CD52), fludarabine and anti-thymocyte globulin (ATG) can eliminate the majority of B and T cells. In addition to alloreactive and autoreactive memory B cells and T cells, other T and B cells are depleted.

b, Immune-mediated destruction of autoreactive cells (graft-versus-autoimmune host T and B cells). Donor T cells recognize alloantigens and destroy residual memory B and T cells. This mechanism is applicable to allo-HCT only. **c**, Deletion of alloreactive and autoreactive T cells in the thymus. If both donor and host cells contribute to haematopoiesis and to the antigen-presenting cell (APC) pool in the thymus, the new T-cell repertoire generated in the recipient thymus is deleted of T cells recognizing both allo- and

autoantigens expressed by haematopoietic cells of both origins. **d**, Induction of anergy and deletion of autoreactive and alloreactive T cells in the periphery. Costimulatory blockade of the CD40–CD154 and CD28–CD80–CD86 pathways in concert with allogeneic BMT can overcome the T-cell barrier to allo-HSC engraftment, which in turn quickly produces donor-specific tolerance and could tolerize cross-reactive autoreactive lymphocytes as well. This mechanism plays a role only in allo-HCT. **e**, Tolerization of peripheral autoreactive/alloreactive T cells by regulatory T cells. CD4⁺CD25⁺ regulatory T cells have been implicated in the maintenance of peripheral tolerance to organ-specific self-antigens as well as alloantigens. The secretion of TGF-β and IL-10 by regulatory T cells has been suggested to mediate this process. Both allo-HCT and auto-HCT could incorporate this mechanism. **f**, Tolerization of autoreactive and alloreactive B cells. In addition to alloantigens, haematopoietic chimaerism could potentially tolerize pre-existing recipient B cells to autoantigens expressed by donor haematopoietic cells. This mechanism plays a role only in allo-HCT.

of immunosuppression in a patient receiving allo-HCT to treat leukaemia²⁰. There are two documented case reports of non-myeloablative allogeneic transplants for Evans' syndrome, a disease characterized by combined autoimmune thrombocytopenia and autoimmune haemolytic anaemia. Complete clinical and immunological remission followed GVHD, in one case elicited by donor lymphocyte infusion^{21,22}.

Although depleting T cells from the HCT product can prevent GVHD, this would be expected to lead to a loss of GVA, similar to the loss of GVL that has been observed with this strategy. Some strategies for separating GVHD from GVL that have been explored in animal models might also favour the achievement of GVA while avoiding GVHD. These include the administration of donor-derived regulatory T cells²³, the direct intra-bone injection of donor bone-marrow stem cells and stromal cells²⁴ and the infusion of alloreactive natural killer cells²⁵. An additional strategy, for which proof of principle has been obtained in pilot clinical trials²⁶, involves the non-myeloablative induction of mixed chimaerism with a regimen that includes extensive T-cell depletion of the initial HCT. When conditioning-induced inflammation has subsided, donor lymphocytes are then transplanted, which mediate a 'lymphohaematopoietic GVH response'. This GVH response eliminates normal and malignant recipient haematopoietic cells but does not migrate to the GVHD target tissues, the skin, gut and liver, and so GVHD does not occur²⁷.

Deletion of alloreactive and autoreactive T cells in the thymus

Allo-HCT and full donor reconstitution leads to intrathymic deletion (or tolerance) of newly developing thymocytes that recognize donor antigens, permitting acceptance of donor organs without immunosuppressive therapy. Mixed haematopoietic chimaerism is a state in which HSCs, and therefore multi-lineage haematopoietic populations, of the recipient and the donor co-exist. Mixed haematopoietic chimaerism has several advantages over full allogeneic chimaerism and is being explored in animal models as an approach to achieving organ-graft tolerance. Mixed haematopoietic chimaerism can be achieved by milder conditioning treatments that do not ablate host haematopoietic cells²⁸, overcoming some of the high treatment-related toxicity that has largely prevented the clinical application of allo-HCT for non-life-threatening diseases. Mixed chimaerism is associated with improved immunocompetence compared with full chimaerism, particularly if the donor and recipient do not share MHC alleles²⁸. Mixed but not full chimaeras contain a life-long source of host-type antigen-presenting cells (APCs) that can most effectively present antigens to T cells that are positively selected in the recipient thymus²⁸ (Fig. 3c).

If both donor and host cells contribute to haematopoiesis, the new T-cell repertoire generated in the recipient thymus is rendered tolerant of antigens expressed by haematopoietic cells of both origins. Haematopoietic cells most efficiently delete thymocytes with high affinity for self-MHC-peptide complexes, thus ensuring tolerance to the donor and host²⁸. Induction of mixed chimaerism prevents the development of autoimmune disease with greater efficiency than auto-HCT does in several animal models. The reason that donor haematopoietic cells can induce tolerance of newly developing autoimmune-prone recipient T cells in these models is not known. It is possible that donor-derived APCs expressing allogeneic MHC-peptide complexes mediate intrathymic deletion of autoreactive host thymocytes with cross-reactive TCRs. However, the observation of similar phenomena in different disease models, using different animal strains, makes this explanation somewhat unsatisfactory.

Genetic susceptibility and resistance to type 1 diabetes are associated with highly polymorphic genes of the MHC. The NOD mouse is homozygous for a unique H-2 haplotype (H-2^{g7}), and the I-A^{g7} allele is an orthologue of the HLA-DQ0302 allele in humans, which confers susceptibility to type 1 diabetes. Although controversial, some studies suggest that MHC class II I-A^{g7} molecules, by binding to self-peptides with low efficiency during thymic selection, permit self-reactive T cells to escape negative selection²⁹. The expression of stable 'protective', cross-reactive allo-MHC-peptide complexes on donor HSC-derived

APCs may result in specific central deletion of host-derived autoreactive T cells.

Studies with I-A^{g7}-restricted islet-cell-reactive TCR-transgenic mice expressing antidiabetogenic MHC class II haplotypes show that protective MHC class II molecules induce deletion of certain diabetogenic TCRs. This negative selection of autoreactive thymocytes is mediated by thymic bone-marrow-derived cells independently from endogenous superantigens³⁰. Intrathymic deletion mediated by 'protective' MHC class II occurs in a transgenic diabetes model involving a diabetogenic TCR from a CD8⁺ T-cell clone³¹, and in NOD mice receiving syngeneic HSC transduced with a protective MHC class II molecule³². In another diabetogenic TCR transgenic model in NOD mice, however, protective MHC class II molecules expressed by F₁ mice tolerize diabetogenic CD4⁺ T cells, not through deletion but through generation of regulatory cell populations³³. In this case, the thymic epithelium expressed the protective MHC molecules. Non-haematopoietic-cell MHC expression is essential for positive selection of regulatory T cells³³, but this mechanism cannot readily explain the tolerance of autoreactive clones in mixed haematopoietic chimaeras, because the donor MHC is expressed only by haematopoietic cells.

Many of the studies showing prevention of autoimmune disease by induction of mixed chimaerism have involved lethal total-body irradiation and reconstitution with allogeneic plus syngeneic HSCs. The formidable and opposing problems of GVHD and failure of engraftment that have been observed in lethally conditioned humans when extensive HLA barriers are crossed make this approach unlikely to have clinical relevance. A clinically relevant protocol for reliably achieving sustained mixed chimaerism across MHC barriers for reversal of autoimmune disease must specifically overcome the host immune barriers to the donor and be devoid of GVHD risk. Over the past 15 years, several such protocols have been successfully developed in animal models, many of which involve *in vivo* monoclonal antibody treatment to eliminate both donor and host T cells *in vivo*. This outcome has been achieved in a small number of patients receiving a non-myeloablative conditioning regimen involving *in vivo* and *ex vivo* T-cell depletion²⁶. However, after lymphoablative conditioning, T cells are regenerated quite slowly by the adult thymus, making non-lymphoablative regimens desirable. More recently, protocols have been developed that substitute a costimulatory blockade for some or all of this T-cell depletion. As discussed below, additional mechanisms must be involved in the tolerization of both pre-existing alloreactive and autoreactive T cells in such protocols.

Anergy and deletion of peripheral auto- and alloreactive T cells

In all normal adults, mature T cells with anti-donor reactivity already exist in the periphery. These cells must be eliminated or inactivated by initial host conditioning to prevent the rejection of the infused donor HSCs. Because the adult thymus is very slow to regenerate T cells after lymphoablative conditioning, it would be desirable to avoid recipient T-cell depletion in regimens for HCT. Instead, blockade of the costimulatory signalling pathways CD40-CD154 and CD28-CD80/CD86 together with allo-HCT can overcome the T-cell barrier to allo-HSC engraftment, which in turn quickly produces donor-specific tolerance^{34,35}. Long-term maintenance of this systemic tolerance is achieved by central deletion of alloreactive T cells, the same as in other mixed-chimaerism models^{34,35}. However, because this approach does not include complete depletion of peripheral recipient T cells before HCT, other mechanisms must allow acceptance of donor bone marrow and tolerize pre-existing recipient T cells (Fig. 3d). The combination of anti-CD154 monoclonal antibodies and allo-HCT tolerizes existing alloreactive CD4⁺ T cells in the periphery by anergy followed by deletion, owing to presentation of donor antigens on APCs in the absence of a CD40-mediated activation signal³⁴. Donor-reactive CD8⁺ T cells also undergo deletion in the periphery, even more rapidly than CD4⁺ T cells in this model³⁶ (Y. Takeuchi *et al.*, unpublished observations), by mechanisms that are currently under investigation.

A non-myeloablative regimen using costimulatory blockade with

sublethal total body irradiation induced tolerance to allogeneic islets transplanted into NOD mice. However, nearly all haematopoietic cells in these animals were derived from donors³⁷. In another model, mixed chimaerism was induced in NOD mice that received MHC-matched allo-HCT. Transplantation of donor islets led to a long-term cure of diabetes, despite the presence of lymphocytic infiltrates in the islet grafts³⁸.

In a recent study, mixed chimaerism was established across extensive MHC barriers in overtly diabetic NOD mice with non-myeloablative conditioning involving anti-CD154 monoclonal antibodies. These mice accepted donor allogeneic islets and were thereby cured of established diabetes, despite significant T-cell reconstitution by the NOD recipients³⁹. Surprisingly, mixed haematopoietic chimaerism reversed destructive autoimmunity, permitting acceptance of syngeneic islet grafts in established diabetic mice, despite the fact that large numbers of pre-existing NOD T cells escaped depletion with anti-T cell monoclonal antibodies³⁹ (B. Nikolic and M. Sykes, unpublished data). Thus, in addition to tolerizing newly developing host T cells intrathymically, unknown peripheral mechanisms must tolerize pre-existing autoreactive NOD memory T cells in this model. Without HCT, non-myeloablative conditioning did not induce tolerance to either syngeneic or allogeneic islets, demonstrating that allo-HSC are crucial for tolerization of diabetogenic and alloreactive T cells.

Tolerization by regulatory T cells

CD4⁺CD25⁺ regulatory T cells are involved in the maintenance of peripheral tolerance to organ-specific self-antigens. Their development is regulated by the forkhead transcription factor Foxp3 (ref. 41), and its genetic deficiency leads to a lack of CD4⁺CD25⁺ regulatory T cells and autoimmune syndromes in humans and mice^{40,41} (see the review by Rudensky and Kronenberg, page 598). Studies of wild-type and Foxp3-deficient mixed chimaerism showed that the gene acts intrinsically to generate regulatory cells only among Foxp3⁺ cells⁴². However, mixed chimaerism involving HSC expressing and lacking Foxp3 allows for the development of Foxp3⁺ regulatory cells that can dominantly inhibit autoimmunity. By this mechanism, mixed chimaerism can correct a variety of autoimmune syndromes associated with intrinsic defects in haematopoietic cells and their ability to generate regulatory cells⁴³ (Fig. 3e).

In contrast to 'haematopoietic-cell-intrinsic' defects in regulatory cell development, it is unlikely that mixed chimaerism can correct autoimmunity caused by thymic epithelial-cell-intrinsic abnormalities. Such defects would include the failure of I-A^{B7} to positively select regulatory T cells³³, a function performed by the thymic epithelium. This would also apply to defects in *Aire* (autoimmune regulator), a gene expressed by medullary thymic epithelial cells that regulates the intrathymic expression of numerous organ-specific self-antigens⁴⁴. *Aire* mutations in mice and humans are associated with autoimmunity^{44,45} owing to the lack of expression of these self-antigens in the thymus. This may result in insufficient central deletion of self-reactive T cells⁴⁴ and a failure to generate specific CD4⁺CD25⁺ regulatory T cells⁴⁶.

Tolerization of autoreactive and alloreactive B cells

Autoantibodies provide diagnostic and prognostic criteria and play a key role in the pathogenesis of numerous autoimmune diseases, leading to considerable interest in B-cell-directed/depleting therapies for autoimmune diseases. In addition to T-cell tolerance, induction of mixed allogeneic chimaerism also results in B-cell tolerance (Fig. 3f). Using an $\alpha(1,3)$ -galactosyltransferase (GalT) wild-type to GalT-deficient mouse HCT model, non-myeloablative mixed chimaerism tolerizes pre-existing B cells producing natural antibody against the carbohydrate Gal epitope expressed by the donor⁴⁷. Furthermore, mixed chimaerism can also tolerize a class-switched IgG anti-Gal memory B-cell response²⁸. The absence of B cells with receptors recognizing Gal in long-term mixed chimaeras suggests a role for clonal deletion/receptor editing in the maintenance of B-cell tolerance²⁸ (see the review by Goodnow *et al.* for more on mechanisms of tolerance, page 590). However, B cells expressing these receptors take time to dis-

appear, suggesting that anergy is the initial tolerance mechanism (T. Kawahara *et al.*, unpublished observations). Thus, mixed chimaerism might tolerize pre-existing recipient B cells recognizing autoantigens that are shared by the donor haematopoietic cells.

Clinical applications of allo-HCT

The substantial risk of morbidity and mortality associated with allo-HCT has generally prevented it from being used to treat patients with autoimmune diseases. Only recently, on the basis of the rationale discussed in the previous sections, encouraging results from animal models, and the development of non-myeloablative conditioning regimens, some investigators have started exploring allo-HCT for the treatment of severe autoimmune diseases. Clinical observations in patients undergoing allo-HCT for the treatment of severe aplastic anaemia or acute leukaemia, who had a coincident autoimmune disease, support this approach. Although most of these patients have achieved remission of the autoimmune disease¹⁵, there are some notable exceptions. Three patients with rheumatoid arthritis, psoriasis and Crohn's disease relapsed despite achieving nearly full donor chimaerism^{48,49}. There may be several reasons for these relapses: a lack of GVHD (GVA); ongoing immune destruction by persisting recipient cells; the use of HLA-identical donors; and the possible existence of similar autoimmune-susceptible genes in the donor.

Allo-HCT has recently been applied as a therapy for Evans' syndrome, a disease characterized by combined autoimmune thrombocytopenia and autoimmune haemolytic anaemia. Complete remission of severe Evans' syndrome was achieved following HLA-identical sibling cord blood transplantation, but the patient died nine months after HCT from acute liver failure of unknown aetiology⁵⁰. Recently, allo-HCT followed by donor leukocyte infusion to enhance GVA led to complete remission of Evans' syndrome for 2 years after transplantation, in association with grade II GVHD²². In seven patients receiving allo-HCT to treat refractory autoimmune cytopenias, one died of treatment-related complications, one with a transient response died of progressive disease, and five had continuous responses⁵¹.

Tolerance induction through HSC chimaerism

Mixed allogeneic chimaerism may induce a reliable and robust form of allo- and autotolerance. If regimens for achieving mixed chimaerism in patients with acceptably low toxicity are developed, this approach might have a role in the treatment of autoimmune diseases. Two centres have published results of their attempts to induce mixed chimaerism for the establishment of allograft tolerance in patients. The Stanford group used a conditioning regimen consisting of total lymphoid irradiation and anti-thymocyte globulin (ATG) followed by transplantation of G-CSF-mobilized HLA-mismatched CD34⁺ cells and a kidney from the same donor followed by post-transplant immunosuppression⁵². Three out of four patients developed transient chimaerism that was lost within the first three months. Two patients were weaned from all immunosuppression after 12 months. They subsequently developed acute rejection episodes and immunosuppressive therapy had to be resumed⁵². Thus, tolerance was not achieved. At Massachusetts General Hospital, we have used combined kidney transplantation and non-myeloablative HCT from HLA-identical related donors to patients with multiple myeloma with renal failure using a regimen involving peri-transplant ATG, cyclophosphamide, thymic irradiation and a short post-transplantation course of cyclosporine^{53,54}. A second protocol involves HLA-haploidentical transplantation to related recipients who do not have malignant disease. Patients are conditioned with cyclophosphamide, anti-CD2 antibody and thymic irradiation followed by combined renal transplantation and HCT from the same HLA-haploidentical donor. Encouraging preliminary results have been obtained (T. Kawai *et al.*, unpublished data).

Another group recently applied a mixed-chimaerism approach to the therapy of severe, refractory rheumatoid arthritis⁵⁵. The patient was conditioned with fludarabine (an antimetabolite, purine antago-

nist), cyclophosphamide, and anti-CD52 (which recognizes T and B cells) and received HLA-matched CD34⁺ HCT, which resulted in mixed chimaerism among T cells (55% donor CD3⁺ cells) and myeloid cells (70% donor CD33⁺ cells). One year after HCT, the patient had suffered from no major infection and had had no acute or chronic GVHD. Remission of rheumatoid arthritis without GVHD or the need for immunomodulatory medications was achieved⁵⁵.

Although these data are promising, the risk of GVHD and the absence of definitive information on the efficacy of auto-HCT, which lacks this risk, make it unclear what role allo-HCT will ultimately play in the treatment of autoimmune diseases. At present, the substantial mortality and morbidity associated with allo-HCT must be carefully weighed against its therapeutic benefits. The careful and proper choice of prospective patients with refractory autoimmune diseases that are resistant to current pharmacological treatments is essential. It is possible that the therapeutic benefit of auto-HCT could be enhanced by the addition of expanded regulatory cell populations, such as CD4⁺CD25⁺ regulatory T cells, or mesenchymal stem cells, which have potent immunosuppressive properties and have recently been reported to ameliorate severe GVHD⁵⁶.

Future directions

The recent extension of the allo-HCT approach from rodents to humans for tolerance induction to alloantigens is highly encouraging. An understanding of the tolerization processes that apply to naive and memory T and B cells will ultimately lead to the development of less toxic, non-myceloablative conditioning regimens. The development of such conditioning regimens will lead to new hope for patients in need of organ transplants and possibly for those with autoimmune diseases. The ultimate role for allo-HCT in the treatment of autoimmune disease will be better defined by results of ongoing auto-HCT trials, and by an improved understanding of disease pathogenesis, the genetic defects underlying different forms of autoimmune disease, and the mechanisms by which HCT may tolerize memory T and B cells.

In particular, the application of allo-HCT followed by transplantation of islets from the same donor has the potential to improve the therapy of patients with type 1 diabetes. Islet transplantation can normalize glucose levels and in theory could prevent some of the complications of diabetes. The immunosuppressive 'Edmonton Protocol' (which includes increased numbers of high-quality islet cells, omission of steroids and immunosuppression with tacrolimus, rapamycin and an anti-interleukin 2 receptor antibody) is a major step forward. Nonetheless, the overall outcome of clinical islet allotransplantation has been somewhat disappointing⁵⁷. After 3–4 years, only 12–25% of transplanted patients were insulin independent⁵⁷, possibly because of the allo- and autoimmune responses combined with complications of the drug therapy. In addition, the life-long immunosuppression needed to achieve allograft acceptance is hard to justify in young type 1 diabetic patients, making tolerance-inducing approaches particularly desirable. Tolerance would permit graft survival without the need for continuous immunosuppressive therapy. Successful tolerance protocols would overcome alloresponses to donor islet alloantigens and could prevent recurrent autoimmunity in the donor islets, as suggested by mixed-chimaerism studies in NOD mice with advanced diabetes³⁹. Either combined HSC and islets from deceased donors or the combined transplantation of embryonic stem cell⁵⁸ or adult stem-cell-derived islets and HSC could ultimately be used to achieve tolerance and cure of type 1 diabetes. In addition, promising approaches to xenotolerance induction, including HCT³⁹, might ultimately allow the use of xenotransplants (grafts from donors of other species such as pigs) that could provide a solution to the shortage of allogeneic islets.

In addition to tolerization of autoreactive and alloreactive cells, HCT might also allow regeneration of tissues that are destroyed by autoimmune disease. This could theoretically occur owing either to regeneration from endogenous stem cells once autoimmunity is reversed, or because of differentiation of HSCs or other cells in the HCT product into non-haematopoietic tissues. Endogenous islet

regeneration has been described following allogeneic HCT in new onset diabetic NOD mice⁶⁰, and transplanted haematopoietic cells have been reported to promote endogenous pancreatic regeneration^{61,62}. Studies in mixed-chimaeric NOD mice with well established diabetes show, however, that endogenous islet regeneration cannot maintain normoglycaemia and that islet transplantation is necessary to correct the disease³⁹. Several groups have reported that HSC can differentiate into hepatic cells, myocardium, kidney cells and other non-haematopoietic cell types, including islet cells^{63,64}, but it is possible that these studies reflect cell fusion rather than differentiation from cells contained in HCT^{65,66}. Although splenocytes have been reported to contain cells that regenerate islets in diabetic NOD mice⁶⁷, other groups have not been successful in reproducing these results⁶⁸. HCT leading to sustained mixed chimaerism in well-established diabetic mice does not achieve similar results³⁹. Thus, the real potential of tissue regeneration from HSC to cure autoimmune disease is controversial and remains an area of intense investigation. ■

- Morton, J. L. & Siegel, B. V. Transplantation of autoimmune potential. Development of antinuclear antibodies in H-2 histocompatible recipients of bone marrow from New Zealand Black mice. *Proc. Natl Acad. Sci. USA* **71**, 216–266 (1974).
- Ikehara, S. *et al.* Organ-specific and systemic autoimmune diseases originate from defects in haematopoietic stem cells. *Proc. Natl Acad. Sci. USA* **87**, 8341–8344 (1990).
- Ikehara, S. *et al.* Rationale for bone marrow transplantation in the treatment of autoimmune diseases. *Proc. Natl Acad. Sci. USA* **82**, 2483–2487 (1985).
- Ikehara, S. Bone marrow transplantation for autoimmune diseases. *Acta Haematol.* **99**, 116–132 (1998).
- Roep, B. O. & Atkinson, M. Animal models have little to teach us about type 1 diabetes: 1. In support of this proposal. *Diabetologia* **47**, 1650–1656 (2004).
- Ikehara, S. *et al.* Long-term observations of autoimmune-prone mice treated for autoimmune disease by allogeneic bone marrow transplantation. *Proc. Natl Acad. Sci. USA* **86**, 3306–3310 (1989).
- Yasumizu, R. *et al.* Treatment of type 1 diabetes mellitus in non-obese diabetic mice by transplantation of allogeneic bone marrow and pancreatic tissue. *Proc. Natl Acad. Sci. USA* **84**, 6555–6557 (1987).
- Pestronk, A., Drachman, D. B., Teoh, R. & Adams, R. N. Combined short-term immunotherapy for experimental autoimmune myasthenia gravis. *Ann. Neurol.* **14**, 235–241 (1983).
- Knaan-Shanzer, S., Houben, P., KinwelBohrè, E. B. M. & van Bekkum, D. W. Remission induction of adjuvant arthritis in rats by total body irradiation and autologous bone marrow transplantation. *Bone Marrow Transplant.* **8**, 333–338 (1992).
- van Gelder, M. & van Bekkum, D. W. Effective treatment of relapsing experimental autoimmune encephalomyelitis with pseudotransplantation of autologous bone marrow transplantation. *Bone Marrow Transplant.* **18**, 1029–1034 (1996).
- van Bekkum, D. W. Experimental basis for the treatment of autoimmune diseases with autologous hematopoietic stem cell transplantation. *Bone Marrow Transplant.* **32**, S37–S39 (2003).
- Burt, R. K., Padilla, J., Begolka, W. S., Canto, M. C. & Miller, S. D. Effect of disease stage on clinical outcome after syngeneic bone marrow transplantation for relapsing experimental autoimmune encephalomyelitis. *Blood* **91**, 2609–2616 (1998).
- Guillaume, T., Rubinstein, D. B. & Symann, M. Immune reconstitution and immunotherapy after autologous hematopoietic stem cell transplantation. *Blood* **92**, 1471–1490 (1998).
- Brodsky, R. A. High-dose cyclophosphamide for aplastic anemia and autoimmunity. *Curr. Opin. Oncol.* **14**, 143–146 (2002).
- Burt, R. K., Traynor, A. E., Craig, R. & Marmont, A. M. The promise of hematopoietic stem cell transplantation for autoimmune diseases. *Bone Marrow Transplant.* **31**, 521–524 (2003).
- Popat, U. & Krance, R. Haematopoietic stem cell transplantation for autoimmune disorders: the American perspective. *Br. J. Haematol.* **126**, 637–649 (2004).
- Hough, R. E., Snowden, J. A. & Wulffraat, N. M. Haematopoietic stem cell transplantation in autoimmune diseases: a European perspective. *Br. J. Haematol.* **128**, 432–459 (2005).
- Tyndall, A. & Saccardi, R. Hematopoietic stem cell transplantation in the treatment of severe autoimmune disease - results from phase I/II studies, prospective randomized trials and future directions. *Clin. Exp. Immunol.* (in press).
- Hinterberger, W., Hinterberger-Fischer, M. & Marmont, A. M. Clinically demonstrable anti-autoimmunity mediated by allogeneic immune cells favorably affects outcome after stem cell transplantation in human autoimmune diseases. *Bone Marrow Transplant.* **30**, 753–759 (2002).
- Slavin, S., Nagler, A., Varadi, G. & Or, R. Graft vs autoimmunity following allogeneic non-myceloablative blood stem cell transplantation in a patient with chronic myelogenous leukemia and severe systemic psoriasis and psoriatic polyarthritis. *Exp. Hematol.* **28**, 853–857 (2000).
- Oyama, Y. *et al.* Allogeneic stem cell transplantation for Evans syndrome. *Bone Marrow Transplant.* **28**, 903–905 (2001).
- Marmont, A. M., Gualandi, F., Van Lint, M. T. & Bacigalupo, A. Refractory Evans' syndrome treated with allogeneic SCT followed by DLI. Demonstration of a graft-versus-autoimmunity effect. *Bone Marrow Transplant.* **31**, 399–402 (2003).
- Edinger, M. *et al.* CD4⁺CD25⁺ regulatory T cells preserve graft-versus-tumor activity while inhibiting graft-versus-host disease after bone marrow transplantation. *Nature Med.* **9**, 1144–1150 (2003).
- Ikehara, S. A novel strategy for allogeneic stem cell transplantation: perfusion method plus intra-bone marrow injection of stem cells. *Exp. Hematol.* **31**, 1142–1146 (2003).
- Ruggeri, L. *et al.* Effectiveness of donor natural killer cell alloreactivity in mismatched hematopoietic transplants. *Science* **295**, 2097–2100 (2002).

26. Spitzer, T. R. *et al.* Nonmyeloablative haploidentical stem-cell transplantation using anti-CD2 monoclonal antibody (MEDI-507)-based conditioning for refractory hematologic malignancies. *Transplantation* **75**, 1748–1751 (2003).
27. Mapara, M. Y. *et al.* Donor lymphocyte infusions mediate superior graft-versus-leukemia effects in mixed compared to fully allogeneic chimeras: a critical role for host antigen-presenting cells. *Blood* **100**, 1903–1909 (2002).
28. Sykes, M. Mixed chimerism and transplant tolerance. *Immunity* **14**, 417–424 (2001).
29. Ridgway, W. M., Fasso, M. & Fathman, C. G. A new look at MHC and autoimmune disease. *Science* **284**, 749–751 (1999).
30. Schmidt, D., Verdaguer, J., Averill, N. & Santamaria, P. A mechanism for the major histocompatibility complex-linked resistance to autoimmunity. *J. Exp. Med.* **186**, 1059–1075 (1997).
31. Morgan, D. J., Nugent, C. T., Raveney, B. J. & Sherman, L. A. In a transgenic model of spontaneous autoimmune diabetes, expression of a protective class II MHC molecule results in thymic deletion of diabetogenic CD8⁺ T cells. *J. Immunol.* **172**, 1000–1008, (2004).
32. Tian, C. *et al.* Prevention of type 1 diabetes by gene therapy. *J. Clin. Invest.* **114**, 969–978 (2004).
33. Luhder, F., Katz, J., Benoist, C. & Mathis, D. Major histocompatibility complex class II molecules can protect from diabetes by positively selecting T cells with additional specificities. *J. Exp. Med.* **187**, 379–387 (1998).
34. Kurtz, J., Wekerle, T. & Sykes, M. Tolerance in mixed chimerism — a role for regulatory cells? *Trends Immunol.* **25**, 518–523 (2004).
35. Wekerle, T. *et al.* Extrathymic T cell deletion and allogeneic stem cell engraftment induced with costimulatory blockade is followed by central T cell tolerance. *J. Exp. Med.* **187**, 2037–2044 (1998).
36. Takeuchi, Y. *et al.* Earlier low-dose TBI or DST overcomes CD8⁺ T-cell-mediated alloresistance to allogeneic marrow in recipients of anti-CD40L. *Am. J. Transplant.* **4**, 31–40 (2004).
37. Seung, E. *et al.* Allogeneic hematopoietic chimerism in mice treated with sublethal myeloablation and anti-CD154 antibody: absence of graft-versus-host disease, induction of skin allograft tolerance, and prevention of recurrent autoimmunity in islet-allografted NOD/Lt mice. *Blood* **95**, 2175–2182 (2000).
38. Wu, T. *et al.* Inducing tolerance to MHC-matched allogeneic islet grafts in diabetic NOD mice by simultaneous islet and bone marrow transplantation under nonirradiative and nonmyeloablative conditioning therapy. *Transplantation* **74**, 22–27 (2002).
39. Nikolic, B. *et al.* Mixed hematopoietic chimerism allows cure of autoimmune diabetes through allogeneic tolerance and reversal of autoimmunity. *Diabetes* **53**, 376–383 (2004).
40. Wraith, D. C., Nicolson, K. S. & Whitley, N. T. Regulatory CD4⁺ T cells and the control of autoimmune disease. *Curr. Opin. Immunol.* **16**, 695–701 (2004).
41. Bennett, C. L. *et al.* The immune dysregulation, polyendocrinopathy, enteropathy, X-linked syndrome (IPEX) is caused by mutations of FOXP3. *Nature Genet.* **27**, 20–21 (2001).
42. Fontenot, J. D., Gavin, M. A. & Rudensky, A. Y. Foxp3 programs the development and function of CD4⁺CD25⁺ regulatory T cells. *Nature Immunol.* **4**, 330–336 (2003).
43. Kukreja, A. *et al.* Multiple immuno-regulatory defects in type-1 diabetes. *J. Clin. Invest.* **109**, 131–140 (2002).
44. Su, M. A. & Anderson, M. S. Aire: an update. *Curr. Opin. Immunol.* **16**, 746–752 (2004).
45. Vogel, A. *et al.* The genetic background of autoimmune polyendocrinopathy-candidiasis-ectodermal dystrophy and its autoimmune disease components. *J. Mol. Med.* **80**, 201–211 (2002).
46. Apostolou, I., Sarukhan, A., Klein, L. & von Boehmer, H. Origin of regulatory T cells with known specificity for antigen. *Nature Immunol.* **3**, 756–763 (2002).
47. Ohdan, H., Yang, Y. G., Shimizu, A., Swenson, K. G., Sykes, M. Mixed chimerism induced without lethal conditioning prevents T cell- and anti-Gal α 1,3Gal-mediated graft rejection. *J. Clin. Invest.* **104**, 281–290 (1999).
48. McKendry, R. J., Huebsch, L. & Leclair, B. Progression of rheumatoid arthritis following bone marrow transplantation. A case report with a 13-year follow up. *Arthritis Rheum.* **39**, 1246–1253 (1996).
49. Snowden, J. A. *et al.* Long-term outcome of autoimmune disease following allogeneic bone marrow transplantation. *Arthritis Rheum.* **41**, 453–459 (1998).
50. Raetz, E., Beatty, P. G. & Adams, R. H. Treatment of severe Evans syndrome with an allogeneic cord blood transplant. *Bone Marrow Transplant.* **20**, 427–429 (1997).
51. Passweg, J. R. *et al.* Haematopoietic stem cell transplantation for refractory autoimmune cytopenia. *Br. J. Haematol.* **125**, 749–755 (2004).
52. Millan, M. T. *et al.* Mixed chimerism and immunosuppressive drug withdrawal after HLA-mismatched kidney and hematopoietic progenitor transplantation. *Transplantation* **73**, 1386–1391 (2002).
53. Buhler, L. H. *et al.* Induction of kidney allograft tolerance after transient lymphohematopoietic chimerism in patients with multiple myeloma and end-stage renal disease. *Transplantation* **74**, 1405–1409 (2002).
54. Spitzer, T. R. *et al.* Combined histocompatible leukocyte antigen-matched donor bone marrow and renal transplantation for multiple myeloma with end stage renal disease: the induction of allograft tolerance through mixed lymphohematopoietic chimerism. *Transplantation* **68**, 480–484 (1999).
55. Burt, R. K. *et al.* Induction of remission of severe and refractory rheumatoid arthritis by allogeneic mixed chimerism. *Arthritis Rheum.* **50**, 2466–2470 (2004).
56. Le Blanc, K. *et al.* Treatment of severe acute graft-versus-host disease with third party haploidentical mesenchymal stem cells. *Lancet* **363**, 1439–1441 (2004).
57. Rother, K. I. & Harlan, D. M. Challenges facing islet transplantation for the treatment of type 1 diabetes mellitus. *J. Clin. Invest.* **114**, 877–878 (2004).
58. Burt, R. K. *et al.* Embryonic stem cells as an alternate marrow donor source: Engraftment without graft-versus-host disease. *J. Exp. Med.* **199**, 895–904 (2004).
59. Lan, P. *et al.* Induction of human T-cell tolerance to porcine xenoantigens through mixed hematopoietic chimerism. *Blood* **103**, 3964–3969 (2004).
60. Zorina, T. D. *et al.* Recovery of the endogenous beta cell function in the NOD model of autoimmune diabetes. *Stem Cells* **21**, 377–388 (2003).
61. Mathews, V. *et al.* Recruitment of bone marrow-derived endothelial cells to sites of pancreatic beta-cell injury. *Diabetes* **53**, 91–98 (2004).
62. Hess, D. *et al.* Bone marrow-derived stem cells initiate pancreatic regeneration. *Nature Biotech.* **21**, 763–770 (2003).
63. Zulewski, H. *et al.* Multipotent nestin-positive stem cells isolated from adult pancreatic islets differentiate ex vivo into pancreatic endocrine, exocrine, and hepatic phenotypes. *Diabetes* **50**, 521–533 (2001).
64. Jiang, Y. *et al.* Pluripotent nature of adult marrow-derived mesenchymal stem cells. *Nature* **418**, 41–49 (2002).
65. Terada, N. *et al.* Bone marrow cells adopt the phenotype of other cells by spontaneous cell fusion. *Nature* **416**, 542–545 (2002).
66. Ying, Q. L., Nichols, J., Evans, E. P., & Smith, A. G. Changing potency by spontaneous fusion. *Nature* **416**, 545–548 (2002).
67. Kodama, S., Kuhlreiber, W., Fujimura, S., Dale, E. A. & Faustman, D. L. Islet regeneration during the reversal of autoimmune diabetes in NOD mice. *Science* **302**, 1223–1227 (2003).
68. Lechner, A. *et al.* No evidence for significant transdifferentiation of bone marrow into pancreatic beta-cells in vivo. *Diabetes* **53**, 616–623 (2004).

Acknowledgements This work was supported by NIH grants, the JDRF Center for Tolerance at Harvard Medical School (MS) and by The American Society of Nephrology/The American Society of Transplantation John Merrill Transplant Scholar Grant Award (BN). We apologize for the omission of many important references that could not be included here because of space constraints. We thank L. Raleza for expert assistance in manuscript preparation.

Competing interests statement The authors declare that they have no competing financial interests.

Simulations of the formation, evolution and clustering of galaxies and quasars

Volker Springel¹, Simon D. M. White¹, Adrian Jenkins², Carlos S. Frenk², Naoki Yoshida³, Liang Gao¹, Julio Navarro⁴, Robert Thacker⁵, Darren Croton¹, John Helly², John A. Peacock⁶, Shaun Cole², Peter Thomas⁷, Hugh Couchman⁵, August Evrard⁸, Jörg Colberg⁹ & Frazer Pearce¹⁰

The cold dark matter model has become the leading theoretical picture for the formation of structure in the Universe. This model, together with the theory of cosmic inflation, makes a clear prediction for the initial conditions for structure formation and predicts that structures grow hierarchically through gravitational instability. Testing this model requires that the precise measurements delivered by galaxy surveys can be compared to robust and equally precise theoretical calculations. Here we present a simulation of the growth of dark matter structure using 2,160³ particles, following them from redshift $z = 127$ to the present in a cube-shaped region 2,230 billion lightyears on a side. In postprocessing, we also follow the formation and evolution of the galaxies and quasars. We show that baryon-induced features in the initial conditions of the Universe are reflected in distorted form in the low-redshift galaxy distribution, an effect that can be used to constrain the nature of dark energy with future generations of observational surveys of galaxies.

Recent large surveys, such as the 2-degree Field Galaxy Redshift Survey (2dFGRS) and the Sloan Digital Sky Survey (SDSS), have characterized, much more accurately than before, both the spatial clustering and the physical properties of low-redshift galaxies. Major ongoing campaigns exploit the new generation of 8-m-class telescopes and the Hubble Space Telescope to acquire data of comparable quality at high redshift. Other surveys target the weak image shear caused by gravitational lensing to extract precise measurements of the distribution of dark matter around galaxies and galaxy clusters. The principal goals of all these surveys are to shed light on how galaxies form, to test the current model for the growth of cosmic structure, and to search for signatures that may clarify the nature of dark matter and dark energy. These goals can be achieved only if the accurate measurements delivered by the surveys can be compared to robust and equally precise theoretical predictions.

Two problems have so far precluded such predictions: (1) accurate estimates of clustering require simulations of extreme dynamic range, encompassing volumes large enough to contain representative populations of rare objects (such as rich galaxy clusters or quasars), yet resolving the formation of individual low-luminosity galaxies; (2) critical aspects of galaxy-formation physics are uncertain and beyond the reach of direct simulation. Such aspects include, for example, the structure of the interstellar medium and its consequences for star formation and for the generation of galactic winds, the ejection and mixing of heavy elements, and active galactic nuclei (AGN) feeding and feedback effects; these must be treated by phenomenological models whose form and parameters are adjusted by trial and error as part of the overall data-modelling process. Here we have developed a framework that combines very large computer simulations of structure formation with post-hoc modelling of galaxy-formation physics to offer a practical solution to these two entwined problems.

During the past two decades, the cold dark matter (CDM) model, augmented with a dark energy field (which may take the form of a cosmological constant, Λ), has been developed into the standard theoretical model for galaxy formation. It assumes that structure grew from weak density fluctuations present in the otherwise homogeneous and rapidly expanding early Universe. These fluctuations are amplified by gravity, eventually turning into the rich structure that we see around us today. Confidence in the validity of this model has been boosted by recent observations. Measurements of the cosmic microwave background (CMB) by the Wilkinson Microwave Anisotropy Probe (WMAP) satellite¹ were combined with the 2dFGRS to confirm the central tenets of the model and to allow an accurate determination of the geometry and matter content of the Universe about 380,000 yr after the Big Bang². The data suggest that the early density fluctuations were a gaussian random field, as predicted by inflationary theory, and that the current energy density is dominated by some form of dark energy. This analysis is supported by the apparent acceleration of the current cosmic expansion inferred from studies of distant supernovae^{3,4}, as well as by the low matter density derived from the baryon fraction of clusters⁵.

While the initial, linear growth of density perturbations can be calculated analytically, the collapse of fluctuations and the subsequent hierarchical build-up of structure is a highly nonlinear process that is accessible only through direct numerical simulation⁶. The dominant mass component, the cold dark matter, is assumed to be made of elementary particles that currently interact only gravitationally, so the collisionless dark matter fluid can be represented by a set of discrete point particles. This representation as an N -body system is a coarse approximation whose fidelity improves as the number of particles in the simulation increases.

The high-resolution simulation described here—dubbed the

¹Max-Planck-Institute for Astrophysics, Karl-Schwarzschild-Strasse 1, 85740 Garching, Germany. ²Institute for Computational Cosmology, Department of Physics, University of Durham, South Road, Durham DH1 3LE, UK. ³Department of Physics, Nagoya University, Chikusa-ku, Nagoya 464-8602, Japan. ⁴Department of Physics & Astronomy, University of Victoria, Victoria, British Columbia V8P 5C2, Canada. ⁵Department of Physics & Astronomy, McMaster University, 1280 Main Street West, Hamilton, Ontario L8S 4M1, Canada. ⁶Institute of Astronomy, University of Edinburgh, Blackford Hill, Edinburgh EH9 3HJ, UK. ⁷Department of Physics & Astronomy, University of Sussex, Falmer, Brighton BN1 9QH, UK. ⁸Department of Physics & Astronomy, University of Michigan, Ann Arbor, Michigan 48109-1120, USA. ⁹Department of Physics & Astronomy, University of Pittsburgh, 3941 O'Hara Street, Pittsburgh, Pennsylvania 15260, USA. ¹⁰Physics and Astronomy Department, University of Nottingham, Nottingham NG7 2RD, UK.

Millennium Simulation because of its size—was carried out by the Virgo Consortium, a collaboration of British, German, Canadian and US astrophysicists. It follows $N = 2,160^3 \approx 1.0078 \times 10^{10}$ particles from redshift $z = 127$ to the present in a cubic region $500h^{-1}$ Mpc on a side, where $1 + z$ is the expansion factor of the Universe relative to the present and h is Hubble's constant in units of $100 \text{ km s}^{-1} \text{ Mpc}^{-1}$. With ten times as many particles as the previous largest computations of this kind^{7–9} (see Supplementary Information), it offers substantially improved spatial and time resolution within a large cosmological volume. Combining this simulation with new techniques for following the formation and evolution of galaxies, we predict the positions, velocities and intrinsic properties of all galaxies brighter than the Small Magellanic Cloud throughout volumes comparable to the largest current surveys. Crucially, this also allows us to establish evolutionary links between objects observed at different epochs. For example, we demonstrate that galaxies with supermassive central black holes can plausibly form early enough in the standard cold dark matter cosmology to host the first known quasars, and that these end up at the centres of rich galaxy clusters today.

Dark matter haloes and galaxies

The mass distribution in a Λ CDM universe has a complex topology, often described as a 'cosmic web'¹⁰. This is visible in Fig. 1 (see also

the corresponding Supplementary Video). The zoomed-out panel at the bottom of the figure reveals a tight network of cold dark matter clusters and filaments of characteristic size $\sim 100h^{-1}$ Mpc. On larger scales, there is little discernible structure and the distribution appears homogeneous and isotropic. Subsequent images zoom in by factors of four onto the region surrounding one of the many rich galaxy clusters. The final image reveals several hundred dark matter substructures, resolved as independent, gravitationally bound objects orbiting within the cluster halo. These substructures are the remnants of dark matter haloes that fell into the cluster at earlier times.

The space density of dark matter haloes at various epochs in the simulation is shown in Fig. 2. At present, there are about 18 million haloes above a detection threshold of 20 particles; 49.6% of all particles are included in these haloes. These statistics provide the most precise determination to date of the mass function of cold dark matter haloes^{11,12}. In the range that is well sampled in our simulation ($z \leq 12$, $M \geq 1.7 \times 10^{10} h^{-1} M_{\odot}$, where $M_{\odot}$ is the solar mass), our results are remarkably well described by the analytic formula proposed by ref. 11 from fits to previous simulations. Theoretical models based on an ellipsoidal excursion set formulation¹³ give a less accurate, but still reasonable, match. However, the commonly used Press–Schechter formula¹⁴ underpredicts the high-mass end of the mass function by up to an order of magnitude. Previous studies of the abundance of rare objects, such as luminous quasars or clusters,

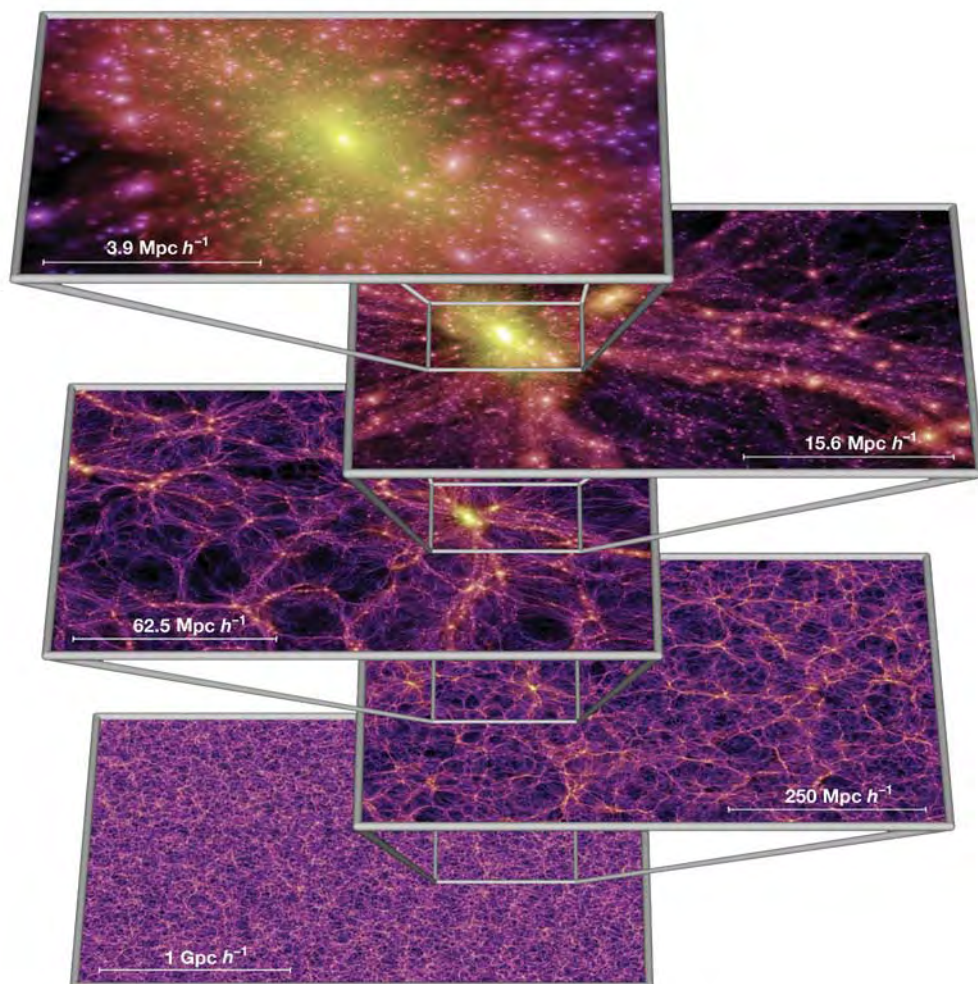


Figure 1 | The dark matter density field on various scales. Each individual image shows the projected dark matter density field in a slab of thickness $15h^{-1}$ Mpc (sliced from the periodic simulation volume at an angle chosen to avoid replicating structures in the lower two images), colour-coded by

density and local dark matter velocity dispersion. The zoom sequence displays consecutive enlargements by factors of four, centred on one of the many galaxy cluster haloes present in the simulation.

based on this formula may contain large errors¹⁵. We return below to the important question of the abundance of quasars at early times.

To track the formation of galaxies and quasars in the simulation, we implement a semi-analytic model to follow gas, star and supermassive black-hole processes within the merger history trees of dark matter haloes and their substructures (see Supplementary Information). The trees contain a total of about 800 million nodes, each corresponding to a dark matter subhalo and its associated galaxies. This methodology allows us to test, during postprocessing, many different phenomenological treatments of gas cooling, star formation, AGN growth, feedback, chemical enrichment and so on. Here, we use an update of models described in refs 16 and 17, which are similar in spirit to previous semi-analytic models^{18–23}; the modelling assumptions and parameters are adjusted by trial and error to fit the observed properties of low-redshift galaxies, primarily their joint luminosity–colour distribution and their distributions of morphology, gas content and central black-hole mass. Our use of a high-resolution simulation, particularly our ability to track the evolution of dark matter substructures, removes much of the uncertainty of the more traditional semi-analytic approaches based on Monte Carlo realizations of merger trees. Our technique provides accurate positions and peculiar velocities for all the model galaxies. It also enables us to follow the evolutionary history of individual objects and thus to investigate the relationship between populations seen at different epochs. It is the ability to establish such evolutionary connections that makes this kind of modelling so powerful for interpreting observational data.

The fate of the first quasars

Quasars are among the most luminous objects in the Universe and can be detected at huge cosmological distances. Their luminosity is thought to be powered by accretion onto a central, supermassive black hole. Bright quasars have now been discovered as far back as redshift $z = 6.43$ (ref. 24), and are believed to harbour central

black holes with a mass a billion times that of the Sun. At redshift $z \approx 6$, their co-moving space density is estimated to be $\sim (2.2 \pm 0.73) \times 10^{-9} h^3 \text{Mpc}^{-3}$ (ref. 25). Whether such extremely rare objects can form at all in a Λ CDM cosmology is unknown.

A volume the size of the Millennium Simulation should contain, on average, just under one quasar at the above space density. Just what sort of object should be associated with these ‘first quasars’ is, however, a matter of debate. In the local Universe, it appears that every bright galaxy hosts a supermassive black hole and there is a remarkably good correlation between the mass of the central black hole and the stellar mass or velocity dispersion of the bulge of the host galaxy²⁶. It would therefore seem natural to assume that, at any epoch, the brightest quasars are always hosted by the largest galaxies. In our simulation, ‘large galaxies’ can be identified in various ways, for example, according to their dark matter halo mass, stellar mass or instantaneous star-formation rate. We have identified the ten ‘largest’ objects defined in these three ways at redshift $z = 6.2$. It turns out that these criteria all select essentially the same objects: the eight largest galaxies by halo mass are identical to the eight largest galaxies by stellar mass; only the ranking differs. Somewhat larger differences are present when galaxies are selected by star-formation rate, but the four first-ranked galaxies are still among the eight identified according to the other two criteria.

In Fig. 3, we illustrate the environment of a ‘first quasar’ candidate in our simulation at $z = 6.2$. The object lies on one of the most prominent dark matter filaments and is surrounded by a large number of other, much fainter galaxies. It has a stellar mass of $6.8 \times 10^{10} h^{-1} M_\odot$, the largest in the entire simulation at $z = 6.2$, a dark matter virial mass of $3.9 \times 10^{12} h^{-1} M_\odot$, and a star-formation rate of $235 M_\odot \text{yr}^{-1}$. In the local Universe, central black-hole masses are typically $\sim 1/1,000$ of the bulge stellar mass²⁷, but in the model we test here these massive early galaxies have black-hole masses in the range 10^8 – $10^9 M_\odot$, significantly larger than low-redshift galaxies of similar stellar mass. To attain the observed luminosities, they must convert infalling mass to radiated energy with a somewhat higher efficiency than the $\sim 0.1c^2$ expected for accretion onto a non-spinning black hole (where c is the speed of light in vacuum).

Within our simulation we can readily address fundamental questions such as: Where are the descendants of the early quasars today? What were their progenitors? By tracking the merging history trees of the host haloes, we find that all our quasar candidates end up today as central galaxies in rich clusters. For example, the object depicted in Fig. 3 lies, today, at the centre of the ninth most massive cluster in the volume, of mass $M = 1.46 \times 10^{15} h^{-1} M_\odot$. The candidate with the largest virial mass at $z = 6.2$ (which has stellar mass $4.7 \times 10^{10} h^{-1} M_\odot$, virial mass $4.85 \times 10^{12} h^{-1} M_\odot$, and star-formation rate $218 M_\odot \text{yr}^{-1}$) ends up in the second most massive cluster, of mass $3.39 \times 10^{15} h^{-1} M_\odot$. Following the merging tree backwards in time, we can trace our quasar candidate back to redshift $z = 16.7$, when its host halo had a mass of only $1.8 \times 10^{10} h^{-1} M_\odot$. At this epoch, it is one of just 18 objects that we identify as collapsed systems with ≥ 20 particles. These results confirm the view that rich galaxy clusters are rather special places. Not only are they the largest virialized structures today, they also lie in the regions where the first structures developed at high redshift. Thus, the best place to search for the oldest stars in the Universe or for the descendants of the first supermassive black holes is at the centres of present-day rich galaxy clusters.

The clustering evolution of dark matter and galaxies

The combination of a large-volume, high-resolution N -body simulation with realistic modelling of galaxies enables us to make precise theoretical predictions for the clustering of galaxies as a function of redshift and intrinsic galaxy properties. These can be compared directly with existing and planned surveys. The two-point correlation function of our model galaxies at redshift $z = 0$ is plotted in Fig. 4 and is compared with a recent measurement from the 2dFGRS

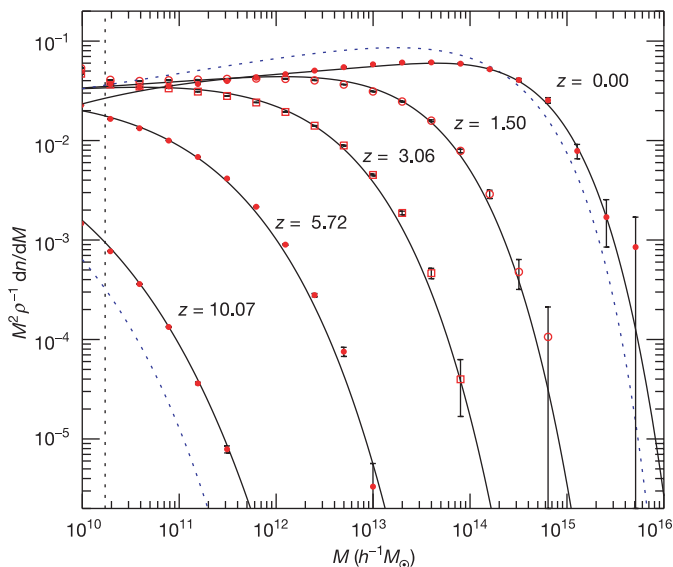


Figure 2 | Differential halo number density as a function of mass and epoch. The function $n(M, z)$ gives the co-moving number density of haloes less massive than M . We plot it as the halo multiplicity function $M^2 \rho^{-1} dn/dM$ (symbols with 1- σ error bars), where ρ is the mean density of the Universe. Groups of particles were found using a friends-of-friends algorithm⁶ with linking length equal to 0.2 of the mean particle separation. The fraction of mass bound to haloes of more than 20 particles (vertical dotted line) grows from 6.42×10^{-4} at $z = 10.07$ to 0.496 at $z = 0$. Solid lines are predictions from an analytic fitting function proposed in previous work¹¹, and the dashed blue lines give the Press–Schechter model¹⁴ at $z = 10.07$ and $z = 0$.

(ref. 28). The prediction is remarkably close to a power law, confirming with much higher precision the results of earlier semi-analytic^{23,29} and hydrodynamic³⁰ simulations. This precision will allow interpretation of the small but measurable deviations from a pure power law found in the most recent data^{31,32}. The simple power-law form contrasts with the more complex behaviour exhibited by the dark matter correlation function, but is really no more than a coincidence. Correlation functions for galaxy samples with different selection criteria or at different redshifts do not, in general, follow power laws.

Although our semi-analytic model was not tuned to match observations of galaxy clustering, it not only produces the excellent overall agreement shown in Fig. 4, but also reproduces the observed dependence of clustering on magnitude and colour in the 2dFGRS and SDSS (refs 33–35), as shown in Fig. 5. The agreement is particularly good for the dependence of clustering on luminosity. The colour dependence of the slope is matched precisely, but the amplitude difference is greater in our model than is observed³⁵. Note that our predictions for galaxy correlations split by colour deviate substantially from power laws. Such predictions can be easily tested against survey data to clarify the physical processes responsible for the observed variations.

In contrast to the near-power-law behaviour of galaxy correlations on small scales, the large-scale clustering pattern may show interesting structure. Coherent oscillations in the primordial plasma give rise to the well-known acoustic peaks in the CMB (refs 2, 36, 37) and also leave an imprint in the linear power spectrum of the dark matter. Detection of these ‘baryon wiggles’ would not only provide an excellent consistency check for the cosmological model, but could also have important practical applications. The characteristic scale of the wiggles provides a ‘standard ruler’ that may be used to constrain the equation of state of the dark energy³⁸. A critical question when designing future surveys is whether these baryon wiggles are present and are detectable in the galaxy distribution, particularly at high redshift.

On large scales and at early times, the mode amplitudes of the dark matter power spectrum grow linearly, roughly in proportion to the cosmological expansion factor. Nonlinear evolution accelerates the growth on small scales when the dimensionless power $\Delta^2(k) = k^3 P(k)/(2\pi^2)$ approaches unity (the power spectrum $P(k)$ measures the variance of density fluctuations on the scale of wavenumber k); this regime can only be studied accurately using numerical simulations. In the Millennium Simulation, we are able to determine the nonlinear power spectrum over a larger range of scales

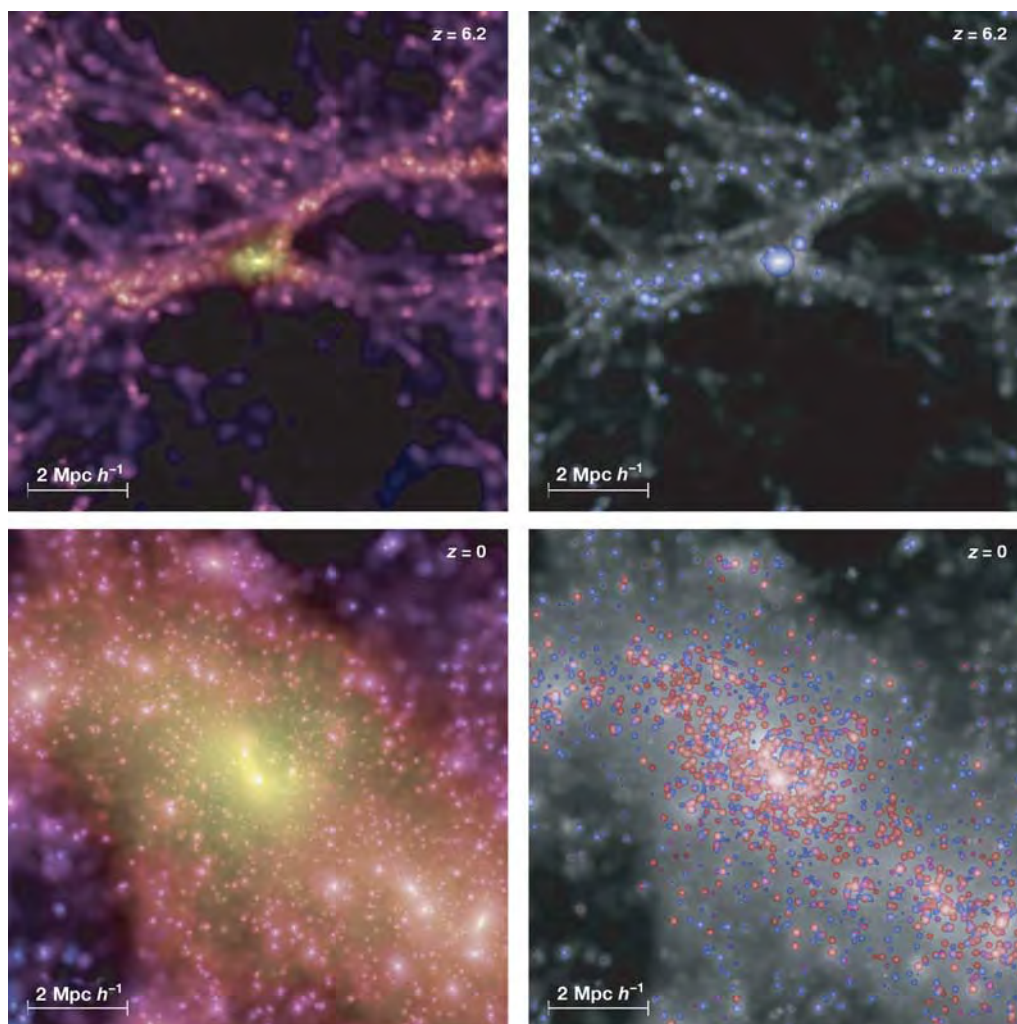


Figure 3 | Environment of a ‘first quasar candidate’ at high and low redshifts. The two panels on the left show the projected dark matter distribution in a cube of co-moving sidelength $10h^{-1}$ Mpc, colour-coded according to density and local dark matter velocity dispersion. The panels on the right show the galaxies of the semi-analytic model overlaid on a greyscale

image of the dark matter density. The volume of the sphere representing each galaxy is proportional to its stellar mass, and the chosen colours encode the restframe stellar B–V colour index. While at $z = 6.2$ (top panels) all galaxies appear blue owing to ongoing star formation, many of the galaxies that have fallen into the rich cluster at $z = 0$ (bottom) have turned red.

than was possible in earlier work³⁹: almost five orders of magnitude in wavenumber k .

At present, the acoustic oscillations in the matter power spectrum are expected to fall in the transition region between linear and nonlinear scales. In Fig. 6, we examine the matter power spectrum in our simulation in the region of the oscillations. Dividing by the smooth power spectrum of a Λ CDM model with no baryons⁴⁰ highlights the baryonic features in the initial power spectrum of the simulation, although there is substantial scatter owing to the small number of large-scale modes. Because linear growth preserves the relative mode amplitudes, we can approximately correct for this scatter by scaling the measured power in each bin by a multiplicative factor based on the initial difference between the actual bin power and the mean power expected in our Λ CDM model. This makes the effects of nonlinear evolution on the baryon oscillations more clearly visible. As Fig. 6 shows, nonlinear evolution not only accelerates growth but also reduces the baryon oscillations: scales near peaks grow slightly more slowly than scales near troughs. This is a consequence of the mode-mode coupling characteristic of nonlinear growth. In spite of these effects, the first two ‘acoustic peaks’ (at $k \approx 0.07$ and $k \approx 0.13 h \text{ Mpc}^{-1}$, respectively) in the dark matter distribution do survive in distorted form until the present day (see Fig. 6f).

Are the baryon wiggles also present in the galaxy distribution? Fig. 6 shows that the answer to this important question is “yes”. The $z = 0$ panel (Fig. 6f) shows the power spectrum for all model galaxies brighter than $M_B = -17$ (M_B is the magnitude in the optical blue waveband). On the largest scales, the galaxy power spectrum has the same shape as that of the dark matter, but with slightly lower amplitude corresponding to an ‘antibias’ of 8%. Samples of brighter galaxies show less antibias, while for the brightest galaxies, the bias becomes slightly positive. Figure 6d, e also shows measurements of the power spectrum of luminous galaxies at redshifts $z = 0.98$ and $z = 3.06$. Galaxies at $z = 0.98$ were selected to have a magnitude $M_B < -19$ in the restframe, whereas galaxies at $z = 3.06$ were selected to have stellar mass larger than $5.83 \times 10^9 h^{-1} M_\odot$, corresponding to a space density of $8 \times 10^{-3} h^3 \text{ Mpc}^{-3}$, similar to that of

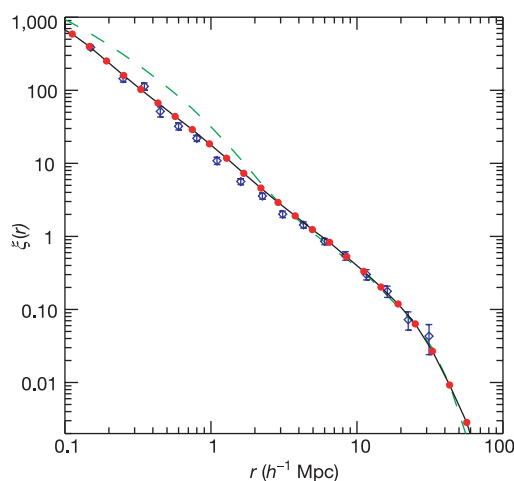


Figure 4 | Galaxy two-point correlation function, $\xi(r)$, at the present epoch as a function of separation r . Red symbols (with vanishingly small Poisson error bars) show measurements for model galaxies brighter than $M_K = -23$, where M_K is the magnitude in the K-band. Data for the large spectroscopic redshift survey 2dFGRS (ref. 28) are shown as blue diamonds together with their $1-\sigma$ error bars. The SDSS (ref. 34) and APM (ref. 31) surveys give similar results. Both for the observational data and for the simulated galaxies, the correlation function is very close to a power law for $r \leq 20 h^{-1} \text{ Mpc}$. By contrast, the correlation function for the dark matter (dashed green line) deviates strongly from a power law.

the Lyman-break galaxies observed at $z \approx 3$ (ref. 41). Signatures of the first two acoustic peaks are clearly visible at both redshifts, even though the density field of the $z = 3$ galaxies is much more strongly biased with respect to the dark matter (by a factor $b = 2.7$, where $b = [P_{\text{gal}}(k)/P_{\text{dm}}(k)]^{1/2}$) than at low redshift. Selecting galaxies by their star-formation rate rather than their stellar mass (above $10.6 M_\odot \text{ yr}^{-1}$ for an equal space density at $z = 3$) produces very similar results.

Our analysis demonstrates conclusively that baryon wiggles should indeed be present in the galaxy distribution out to redshift $z = 3$. This has been assumed, but not justified, in recent proposals to use evolution of the large-scale galaxy distribution to constrain the nature of the dark energy. To establish whether the baryon oscillations can be measured in practice with the requisite accuracy will require detailed modelling of the selection criteria of an actual survey, and a thorough understanding of the systematic effects that will inevitably be present in real data. These issues can only be properly addressed by means of specially designed mock

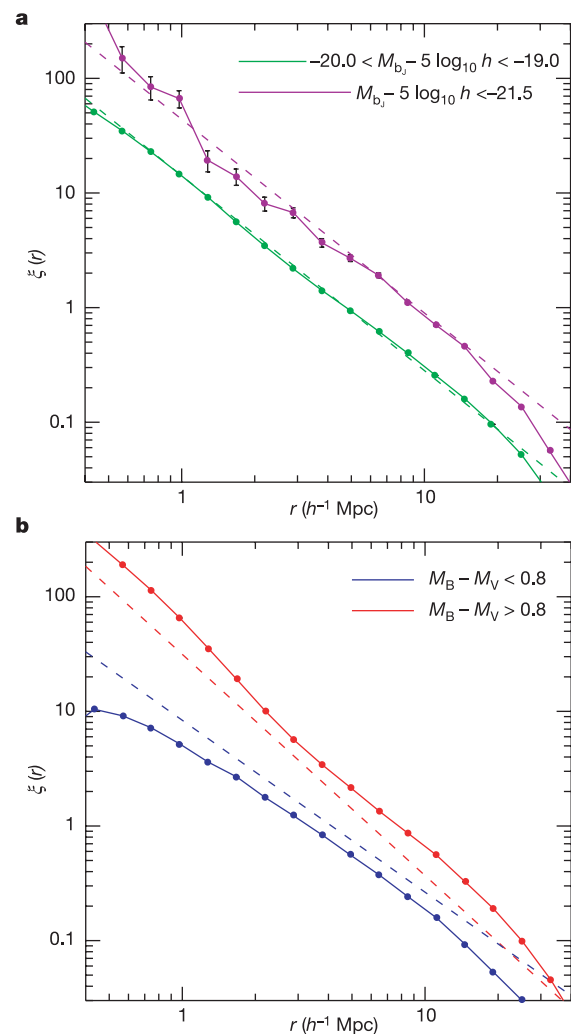


Figure 5 | Galaxy clustering as a function of luminosity and colour. **a**, the two-point correlation function of our galaxy catalogue at $z = 0$ split by luminosity in the b_J -band filter (symbols with $1-\sigma$ error bars). Brighter galaxies are more strongly clustered, in quantitative agreement with observations³³ (dashed lines). Splitting galaxies according to colour (**b**), we find that red galaxies are more strongly clustered with a steeper correlation slope than blue galaxies. Observations³⁵ (dashed lines) show a similar trend, although the difference in clustering amplitude is smaller than in this particular semi-analytic model.

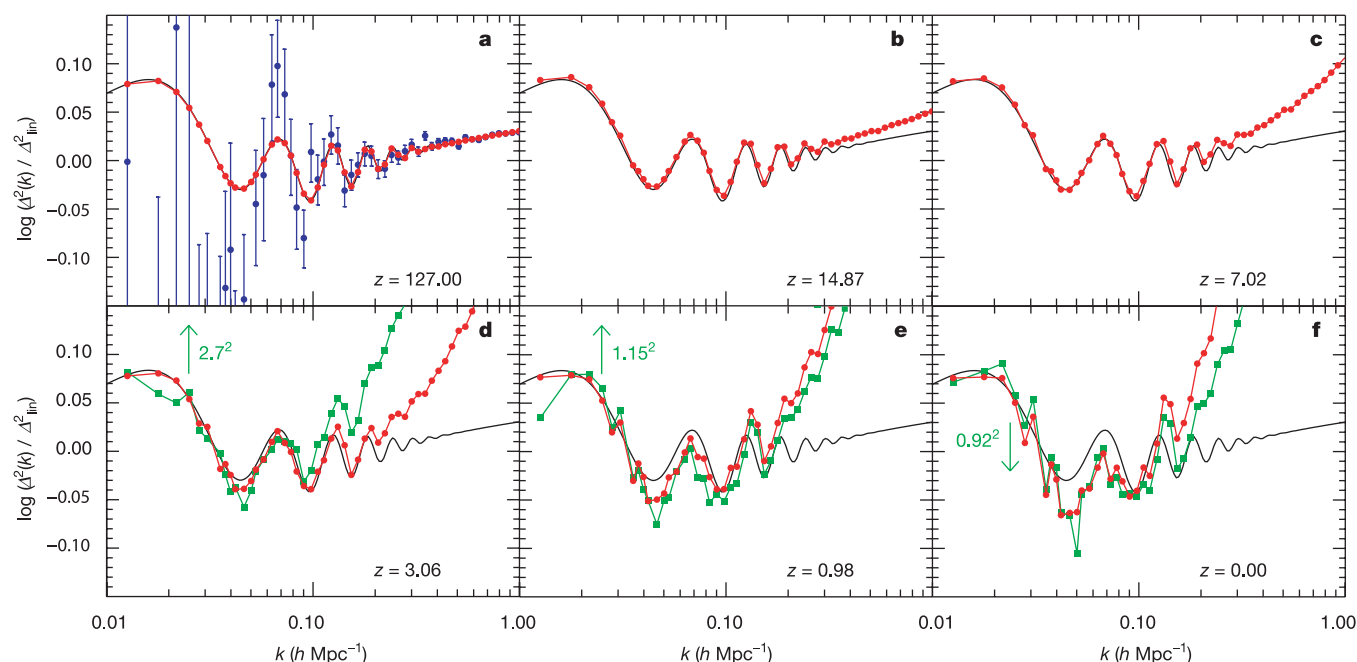


Figure 6 | Power spectra of the dark matter and galaxy distributions in the baryon oscillation region. **a–f**, All measurements have been divided by a linearly evolved, CDM-only power spectrum Δ_{lin}^2 (ref. 40). Red circles show the dark matter, and green squares the galaxies. Blue symbols give the actual realization of the initial fluctuations in our simulation, which scatters around the mean input power (black lines) owing to the finite number of modes (error bars give the 1- σ scatter around the mean power in each bin). Because linear growth preserves relative mode amplitudes, we correct the power in each bin to the expected input power and apply

these scaling factors at all other times. **d**, At $z = 3.06$, galaxies with stellar mass above $5.83 \times 10^9 h^{-1} M_{\odot}$ and space-density of $8 \times 10^{-3} h^3 \text{Mpc}^{-3}$ were selected. Their large-scale density field is biased by a factor $b = 2.7$ with respect to the dark matter (the galaxy measurement has been divided by b^2). **f**, At $z = 0$, galaxies brighter than $M_B = -17$ and a space density higher by a factor of ~ 7.2 were selected. They exhibit a slight antibias of $b = 0.92$. **e**, The corresponding numbers for $z = 0.98$ are $M_B = -19$ and $b = 1.15$.

catalogues constructed from realistic simulations. We plan to construct suitable mock catalogues from the Millennium Simulation and make them publicly available. Our provisional conclusion, however, is that the next generation of galaxy surveys offers excellent prospects for constraining the equation of state of the dark energy.

N -body simulations of CDM universes are now of such size and quality that realistic modelling of galaxy formation in volumes matched to modern surveys has become possible. Detailed studies of galaxy and AGN evolution exploiting the unique data set of the Millennium Simulation therefore make stringent new tests of the theory of hierarchical galaxy formation possible. Using the simulation, we demonstrate that quasars can plausibly form sufficiently early in a Λ CDM universe to be compatible with observation, that their progenitors were already massive by $z \approx 16$, and that their $z = 0$ descendants lie at the centres of cD galaxies in rich galaxy clusters. Interesting tests of our predictions will become possible if observations of the black-hole demographics can be extended to high redshift, allowing, for example, a measurement of the evolution of the relationship between supermassive black-hole masses and the velocity dispersion of their host stellar bulges.

We have also demonstrated that a power-law galaxy autocorrelation function can arise naturally in a Λ CDM universe, but that this behaviour has no simple physical cause and is merely a coincidence. Galaxy surveys will soon reach sufficient statistical power to measure precise deviations from power laws for galaxy subsamples, and we expect that comparisons of the kind we have illustrated will lead to tight constraints on the physical processes included in the galaxy-formation modelling. Finally, we have demonstrated that the baryon-induced oscillations recently detected in the CMB power spectrum should survive in distorted form not only in the nonlinear dark matter power spectrum at low redshift, but also in the power spectra

of realistically selected galaxy samples at $0 < z < 3$. Present galaxy surveys are marginally able to detect the baryonic features at low redshifts^{42,43}. If future surveys improve on this and reach sufficient volume and galaxy density also at high redshift, then precision measurements of galaxy clustering will shed light on one of the most puzzling components of the Universe: the elusive dark energy field.

METHODS

The Millennium Simulation was carried out with a customized version of the GADGET2 (ref. 44) code, using the 'TreePM' method⁴⁵ for evaluating gravitational forces. This is a combination of a hierarchical multipole expansion, or 'tree' algorithm⁴⁶, and a classical Fourier-transform particle-mesh method⁴⁷. The calculation was performed on 512 processors of an IBM p690 parallel computer at the Computing Centre of the Max-Planck Society in Garching, Germany. It used almost all of the 1 terabyte of physically distributed memory available. It required about 350,000 processor hours of CPU time, or 28 days of wall-clock time. The mean sustained floating-point performance (as measured by hardware counters) was about 0.2 teraflops, so the total number of floating-point operations carried out was of the order of 5×10^{17} .

Parameters and initial conditions. The cosmological parameters of our Λ CDM-simulation are: $\Omega_m = \Omega_{\text{dm}} + \Omega_b = 0.25$, $\Omega_b = 0.045$, $h = 0.73$, $\Omega_{\Lambda} = 0.75$, $n = 1$, and $\sigma_8 = 0.9$. Here Ω_m denotes the total matter density in units of the critical density for closure, $\rho_{\text{crit}} = 3H_0^2/(8\pi G)$, where G is Newton's gravitational constant. Similarly, Ω_b and Ω_{Λ} denote the densities of baryons and dark energy at the present day. The Hubble constant is parameterized as $H_0 = 100 h \text{ km s}^{-1} \text{ Mpc}^{-1}$, while σ_8 is the root-mean-square (r.m.s.) linear mass fluctuation within a sphere of radius $8h^{-1} \text{ Mpc}$ extrapolated to $z = 0$. Our adopted parameter values are consistent with a combined analysis of the 2dFGRS (ref. 48) and first-year WMAP data².

The simulation volume is a periodic box of size $500h^{-1} \text{ Mpc}$ and individual particles have a mass of $8.6 \times 10^8 h^{-1} M_{\odot}$. This volume is large enough to include interesting rare objects, but still small enough that the haloes of all luminous galaxies brighter than $0.1L_*$ (where L_* is the characteristic luminosity of galaxies)

are resolved with at least 100 particles. At present, the richest clusters of galaxies contain about three million particles. The gravitational force law is softened isotropically on a co-moving scale of $5h^{-1}$ kpc (Plummer-equivalent), which may be taken as the spatial resolution limit of the calculation. Thus, our simulation achieves a dynamic range of 10^5 in three dimensions, and this resolution is available everywhere in the simulation volume.

Initial conditions were laid down by perturbing a homogeneous, 'glass-like', particle distribution⁴⁹ with a realization of a gaussian random field with the Λ CDM linear power spectrum as given by the Boltzmann code CMBFAST⁵⁰. The displacement field in Fourier space was constructed using the Zel'dovich approximation, with the amplitude of each random phase mode drawn from a Rayleigh distribution. The simulation started at redshift $z = 127$ and was evolved to the present using a leapfrog integration scheme with individual and adaptive timesteps, with up to 11,000 timesteps for individual particles. We stored the full particle data at 64 output times, each of size 300 gigabytes, giving a raw data volume of nearly 20 terabytes. This allowed the construction of finely resolved hierarchical merging trees for tens of millions of haloes and for the subhaloes that survive within them. A galaxy catalogue for the full simulation, typically containing $\sim 2 \times 10^7$ galaxies at $z = 0$ together with their full histories, can then be built for any desired semi-analytic model in a few hours on a small workstation cluster.

Galaxy-formation modelling. The semi-analytic model itself can be viewed as a simplified simulation of the galaxy-formation process, where the star formation and its regulation by feedback processes is parameterized in terms of simple analytic physical models. These models take the form of differential equations for the time evolution of the galaxies that populate each hierarchical merging tree. In brief, these equations describe radiative cooling of gas, star formation, growth of supermassive black holes, feedback processes by supernovae and AGN, and effects due to a re-ionising ultraviolet background. The morphological transformation of galaxies and the process of metal enrichment are modelled as well. To make direct contact with observational data, we apply modern population synthesis models to predict spectra and magnitudes for the stellar light emitted by galaxies, also including simplified models for dust obscuration. In this way we can match the passbands commonly used in observations.

The basic elements of galaxy-formation modelling follow previous studies^{16,18–23} (see also Supplementary Information), but we also use new approaches in a number of areas. First, and of substantial importance, is our tracking of dark matter substructure. This we carry out consistently and with unprecedented resolution throughout our large cosmological volume, allowing an accurate determination of the orbits of galaxies within larger structures, as well as robust estimates of the survival time of structures infalling into larger objects. Also, we use dark matter halo properties, such as angular momentum or density profile, to determine directly sizes of galactic disks and their rotation curves. Second, we use a new model for the build-up of a population of supermassive black holes in the Universe. To this end we extend the quasar model developed in previous work¹⁷ with a 'radio mode', which describes the feedback activity of central AGN in groups and clusters of galaxies. Although largely unimportant for the cumulative growth of the total black-hole mass density in the Universe, our results show that the radio mode becomes important at low redshift, where it has a strong impact on cluster cooling flows, reddening and reducing the brightness of central cluster galaxies. This shapes the bright end of the galaxy luminosity function, bringing our predictions into good agreement with observation.

Received 4 January; accepted 25 March 2005.

- Bennett, C. L. *et al.* First-year Wilkinson Microwave Anisotropy Probe (WMAP) observations: Preliminary maps and basic results. *Astrophys. J. Suppl.* **148**, 1–27 (2003).
- Spergel, D. N. *et al.* First-year Wilkinson Microwave Anisotropy Probe (WMAP) observations: Determination of cosmological parameters. *Astrophys. J. Suppl.* **148**, 175–194 (2003).
- Riess, A. G. *et al.* Observational evidence from supernovae for an accelerating universe and a cosmological constant. *Astron. J.* **116**, 1009–1038 (1998).
- Perlmutter, S. *et al.* Measurements of omega and lambda from 42 high-redshift supernovae. *Astrophys. J.* **517**, 565–586 (1999).
- White, S. D. M., Navarro, J. F., Evrard, A. E. & Frenk, C. S. The baryon content of galaxy clusters: a challenge to cosmological orthodoxy. *Nature* **366**, 429–433 (1993).
- Davis, M., Efstathiou, G., Frenk, C. S. & White, S. D. M. The evolution of large-scale structure in a universe dominated by cold dark matter. *Astrophys. J.* **292**, 371–394 (1985).
- Colberg, J. M. *et al.* Clustering of galaxy clusters in cold dark matter universes. *Mon. Not. R. Astron. Soc.* **319**, 209–214 (2000).
- Evrard, A. E. *et al.* Galaxy clusters in Hubble volume simulations: Cosmological

- constraints from sky survey populations. *Astrophys. J.* **573**, 7–36 (2002).
- Wambsganss, J., Bode, P. & Ostriker, J. P. Giant arc statistics in concord with a concordance lambda cold dark matter universe. *Astrophys. J.* **606**, L93–L96 (2004).
- Bond, J. R., Kofman, L. & Pogosyan, D. How filaments of galaxies are woven into the cosmic web. *Nature* **380**, 603–606 (1996).
- Jenkins, A. *et al.* The mass function of dark matter haloes. *Mon. Not. R. Astron. Soc.* **321**, 372–384 (2001).
- Reed, D. *et al.* Evolution of the mass function of dark matter haloes. *Mon. Not. R. Astron. Soc.* **346**, 565–572 (2003).
- Sheth, R. K. & Tormen, G. An excursion set model of hierarchical clustering: ellipsoidal collapse and the moving barrier. *Mon. Not. R. Astron. Soc.* **329**, 61–75 (2002).
- Press, W. H. & Schechter, P. Formation of galaxies and clusters of galaxies by self-similar gravitational condensation. *Astrophys. J.* **187**, 425–438 (1974).
- Efstathiou, G. & Rees, M. J. High-redshift quasars in the Cold Dark Matter cosmogony. *Mon. Not. R. Astron. Soc.* **230**, 5–11 (1988).
- Springel, V., White, S. D. M., Tormen, G. & Kauffmann, G. Populating a cluster of galaxies. – I. Results at $z = 0$. *Mon. Not. R. Astron. Soc.* **328**, 726–750 (2001).
- Kauffmann, G. & Haehnelt, M. A unified model for the evolution of galaxies and quasars. *Mon. Not. R. Astron. Soc.* **311**, 576–588 (2000).
- White, S. D. M. & Frenk, C. S. Galaxy formation through hierarchical clustering. *Astrophys. J.* **379**, 52–79 (1991).
- Kauffmann, G., White, S. D. M. & Guiderdoni, B. The formation and evolution of galaxies within merging dark matter haloes. *Mon. Not. R. Astron. Soc.* **264**, 201–218 (1993).
- Cole, S., Aragon-Salamanca, A., Frenk, C. S., Navarro, J. F. & Zepf, S. E. A recipe for galaxy formation. *Mon. Not. R. Astron. Soc.* **271**, 781–806 (1994).
- Baugh, C. M., Cole, S. & Frenk, C. S. Evolution of the Hubble sequence in hierarchical models for galaxy formation. *Mon. Not. R. Astron. Soc.* **283**, 1361–1378 (1996).
- Somerville, R. S. & Primack, J. R. Semi-analytic modelling of galaxy formation: the local Universe. *Mon. Not. R. Astron. Soc.* **310**, 1087–1110 (1999).
- Kauffmann, G., Colberg, J. M., Diaferio, A. & White, S. D. M. Clustering of galaxies in a hierarchical universe. – I. Methods and results at $z = 0$. *Mon. Not. R. Astron. Soc.* **303**, 188–206 (1999).
- Fan, X. *et al.* A survey of $z > 5.7$ quasars in the Sloan Digital Sky Survey. II. Discovery of three additional quasars at $z > 6$. *Astron. J.* **125**, 1649–1659 (2003).
- Fan, X. *et al.* A survey of $z > 5.7$ quasars in the Sloan Digital Sky Survey. III. Discovery of five additional quasars. *Astron. J.* **128**, 515–522 (2004).
- Tremaine, S. *et al.* The slope of the black hole mass versus velocity dispersion correlation. *Astrophys. J.* **574**, 740–753 (2002).
- Merritt, D. & Ferrarese, L. Black hole demographics from the $M_{\text{BH}}-\sigma$ relation. *Mon. Not. R. Astron. Soc.* **320**, L30–L34 (2001).
- Hawkins, E. *et al.* The 2dF Galaxy Redshift Survey: correlation functions, peculiar velocities and the matter density of the universe. *Mon. Not. R. Astron. Soc.* **346**, 78–96 (2003).
- Benson, A. J., Cole, S., Frenk, C. S., Baugh, C. M. & Lacey, C. G. The nature of galaxy bias and clustering. *Mon. Not. R. Astron. Soc.* **311**, 793–808 (2000).
- Weinberg, D. H., Davé, R., Katz, N. & Hernquist, L. Galaxy clustering and galaxy bias in a Λ CDM universe. *Astrophys. J.* **601**, 1–21 (2004).
- Padilla, N. D. & Baugh, C. M. The power spectrum of galaxy clustering in the APM survey. *Mon. Not. R. Astron. Soc.* **343**, 796–812 (2003).
- Zehavi, I. *et al.* On departures from a power law in the galaxy correlation function. *Astrophys. J.* **608**, 16–24 (2004).
- Norberg, P. *et al.* The 2dF Galaxy Redshift Survey: luminosity dependence of galaxy clustering. *Mon. Not. R. Astron. Soc.* **328**, 64–70 (2001).
- Zehavi, I. *et al.* Galaxy clustering in early Sloan Digital Sky Survey redshift data. *Astrophys. J.* **571**, 172–190 (2002).
- Madgwick, D. S. *et al.* The 2dF Galaxy Redshift Survey: galaxy clustering per spectral type. *Mon. Not. R. Astron. Soc.* **344**, 847–856 (2003).
- de Bernardis, P. *et al.* A flat Universe from high-resolution maps of the cosmic microwave background radiation. *Nature* **404**, 955–959 (2000).
- Mauskopf, P. D. *et al.* Measurement of a peak in the Cosmic Microwave Background power spectrum from the North American test flight of Boomerang. *Astrophys. J.* **536**, L59–L62 (2000).
- Blake, C. & Glazebrook, K. Probing dark energy using baryonic oscillations in the galaxy power spectrum as a cosmological ruler. *Astrophys. J.* **594**, 665–673 (2003).
- Jenkins, A. *et al.* Evolution of structure in cold dark matter universes. *Astrophys. J.* **499**, 20–40 (1998).
- Bardeen, J. M., Bond, J. R., Kaiser, N. & Szalay, A. S. The statistics of peaks of Gaussian random fields. *Astrophys. J.* **304**, 15–61 (1986).
- Adelberger, K. L. *et al.* A counts-in-cells analysis of Lyman-break galaxies at redshift $Z = 3$. *Astrophys. J.* **505**, 18–24 (1998).
- Cole, S. *et al.* The 2dF Galaxy Redshift Survey: Power-spectrum analysis of the final dataset and cosmological implications. *Mon. Not. R. Astron. Soc.* (submitted); preprint at (<http://xxx.lanl.gov/astro-ph/0501174>) (2005).
- Eisenstein, D. J. *et al.* Detection of the baryon acoustic peak in the large-scale correlation function of SDSS luminous red galaxies. *Astrophys. J.* (submitted);

- preprint at (<http://xxx.lanl.gov/astro-ph/0501171>) (2005).
44. Springel, V., Yoshida, N. & White, S. D. M. GADGET: a code for collisionless and gasdynamical cosmological simulations. *N. Astron.* **6**, 79–117 (2001).
 45. Xu, G. A new parallel n-body gravity solver: TPM. *Astrophys. J. Suppl.* **98**, 355–366 (1995).
 46. Barnes, J. & Hut, P. A hierarchical $O(N \log N)$ force-calculation algorithm. *Nature* **324**, 446–449 (1986).
 47. Hockney, R. W. & Eastwood, J. W. *Computer Simulation Using Particles* Ch. 5 (McGraw-Hill, New York, 1981).
 48. Colless, M. *et al.* The 2dF Galaxy Redshift Survey: spectra and redshifts. *Mon. Not. R. Astron. Soc.* **328**, 1039–1063 (2001).
 49. White, S. D. M. in *Cosmology and Large-Scale Structure* (eds Schaefer, R., Silk, J., Spiro, M. & Zinn-Justin, J.) Ch. 8 (Elsevier, Dordrecht, 1996).
 50. Seljak, U. & Zaldarriaga, M. A line-of-sight integration approach to Cosmic Microwave Background anisotropies. *Astrophys. J.* **469**, 437–444 (1996).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements The computations reported here were performed at the Rechenzentrum der Max-Planck-Gesellschaft in Garching, Germany.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to V.S. (vspringel@mpa-garching.mpg.de).

Interchromosomal associations between alternatively expressed loci

Charalampos G. Spilianakis¹, Maria D. Lalioti^{2*}, Terrence Town^{1*}, Gap Ryol Lee¹ & Richard A. Flavell^{1,3}

The T-helper-cell 1 and 2 (T_H1 and T_H2) pathways, defined by cytokines interferon- γ (IFN- γ) and interleukin-4 (IL-4), respectively, comprise two alternative CD4⁺ T-cell fates, with functional consequences for the host immune system. These cytokine genes are encoded on different chromosomes. The recently described T_H2 locus control region (LCR) coordinately regulates the T_H2 cytokine genes by participating in a complex between the LCR and promoters of the cytokine genes *Il4*, *Il5* and *Il13*. Although they are spread over 120 kilobases, these elements are closely juxtaposed in the nucleus in a poised chromatin conformation. In addition to these intrachromosomal interactions, we now describe interchromosomal interactions between the promoter region of the IFN- γ gene on chromosome 10 and the regulatory regions of the T_H2 cytokine locus on chromosome 11. DNase I hypersensitive sites that comprise the T_H2 LCR developmentally regulate these interchromosomal interactions. Furthermore, there seems to be a cell-type-specific dynamic interaction between interacting chromatin partners whereby interchromosomal interactions are apparently lost in favour of intrachromosomal ones upon gene activation. Thus, we provide an example of eukaryotic genes located on separate chromosomes associating physically in the nucleus via interactions that may have a function in coordinating gene expression.

The subnuclear localization of chromatin is not random, and specific genetic loci or whole chromosomes reside in specific locations within the nucleus¹, where each chromosome is localized in a limited and specific space, called a 'territory', and DNA sequences within a chromosome are organized into euchromatin or heterochromatin. In some cases, large chromosomal loops containing active genes extend outside of the defined chromosomal territories². As cellular differentiation proceeds, changes in transcriptional activity are often coupled with changes in subnuclear localization of chromosomes. Silent genes in developing B and T cells are repositioned in the nucleus at pericentromeric heterochromatin^{3–5}. Transcriptional regulatory elements such as locus control regions, enhancers or insulators act by repositioning specific genetic loci to regions with active or silent transcription⁶. Furthermore, sequence-specific DNA-binding proteins may confer their action by directly repositioning these loci to relevant chromatin compartments^{7–12}.

It has been proposed that transcription and RNA-processing events occur in the space between the chromosome territories in an area called the interchromosome domain compartment¹. The higher order of chromatin organization and its links with transcription have been extensively studied at different levels, such as the chromosomal territories or large-scale structures greater in size than the 30-nm chromatin fibre^{13,14}. Although all studies so far have focused on the inter-relationship between transcriptional activation and subnuclear localization, there has been no evidence for direct physical interaction between inter-related genetic loci localized on different chromosomes. The recently developed chromosome conformation capture technique (3C)^{15–18} provides a powerful tool to study interchromosomal interactions with high accuracy in molecular terms.

We recently described large-scale intrachromosomal interactions during the differentiation of naive CD4⁺ T cells to T_H1 or T_H2 cells.

In direct correlation with the expression profiles of the T_H2 cytokines, several DNase I hypersensitive sites have been characterized for the T_H2 cytokine gene locus on mouse chromosome 11 (refs 19–25). Similarly, the expression of the IFN- γ gene (*Ifng*) located on mouse chromosome 10 is regulated by two conserved regions: *Ifng* CNS1 located 5 kilobases (kb) upstream and *Ifng* CNS2 located 18 kb downstream of the initiation codon of the murine *Ifng* gene^{25,26}. We now report that the *Ifng* gene located on mouse chromosome 10 physically interacts with the *Il5* cytokine promoter, *Rad50* promoter and RHS6 DNase I hypersensitive site of the T_H2 LCR, all located in the T_H2 cytokine gene locus on mouse chromosome 11. The interactions are extremely strong in naive CD4⁺ T cells and are subsequently greatly reduced after the differentiation of naive T cells to effector T_H1 or T_H2 cells. In contrast, non-T-cell types do not exhibit these interactions. Concomitant with the loss of these interchromosomal interactions, in T_H1 cells the *Ifng* gene region instead strengthens intrachromosomal interactions with the *Ifng* CNS2 region located downstream of the *Ifng* gene.

We suggest that there are dynamic intra- and interchromosomal interactions between specific genetic loci that regulate transcriptional initiation or silencing of these loci. We also suggest that one of the functions of LCRs in addition to regulating the expression of adjacent genes is to participate in developmentally regulated interchromosomal interaction of related genes. It is likely that interchromosomal association of genes that are coordinately regulated will be a common feature of gene organization in the nucleus of most higher organisms.

Interchromosomal interactions between *Ifng* and T_H2 loci. In cells of the T-cell lineage and other cell types we have recently reported that specific intrachromosomal interactions occur in the T_H2 cytokine gene locus containing the cytokine genes *Il4*, *Il5* and *Il13*. The promoters of the cytokine genes are located in close proximity to the

¹Section of Immunobiology, and ²Department of Genetics, Yale University School of Medicine, New Haven, Connecticut 06520, USA. ³The Howard Hughes Medical Institute, New Haven, Connecticut 06520, USA.

*These authors contributed equally to this work.

T_H2 LCR in the nucleus of certain cell types, probably functioning to facilitate coordinated transcriptional regulation of cytokine-encoding genes¹⁸. In an attempt to unravel the regulatory effects of the T_H2 LCR in other genes, we used the 3C technique to identify whether there are any physical interchromosomal interactions of the T_H2 locus with the murine *Ifng* locus, the signature cytokine of the alternative T_H1 cell lineage. The organization of the loci and the fragments we used in the 3C analysis are depicted in Fig. 1a–e. As a control locus for the 3C analysis, we used the murine *Gm-csf* gene (also called *Csf2*), which is expressed non-selectively in both T_H1 and T_H2 cells, and is located on the same mouse chromosome (chromosome 11) as the T_H2 locus, about 620 kb upstream of the *Il4* gene. We used the equations in Fig. 1f, g in order to obtain the crosslinking frequency between two fragments. As described previously^{16,18}, we corrected the values of the polymerase chain

reaction (PCR) signals we obtained for a given cell type with those obtained for adjacent fragments of the *Gapd* or the β -globin locus, to control for chance interactions (see Supplementary Methods). Clearly, these are highly conservative background controls to choose, because the physical distance between these control fragments is so small as to render them in complete linkage disequilibrium, whereas for DNA segments located on different chromosomes, random association of fragments would occur much less frequently. Thus, we probably underestimate interchromosomal interactions.

We performed the 3C analysis for the *Ifng* locus and the T_H2 cytokine gene locus. Surprisingly, strong and specific interactions were detected between the *Ifng* gene and three fragments of the T_H2 cytokine locus, covering the constitutive DNase I hypersensitive site RHS6 of the T_H2 LCR (fragment H), the *Rad50* promoter (fragment

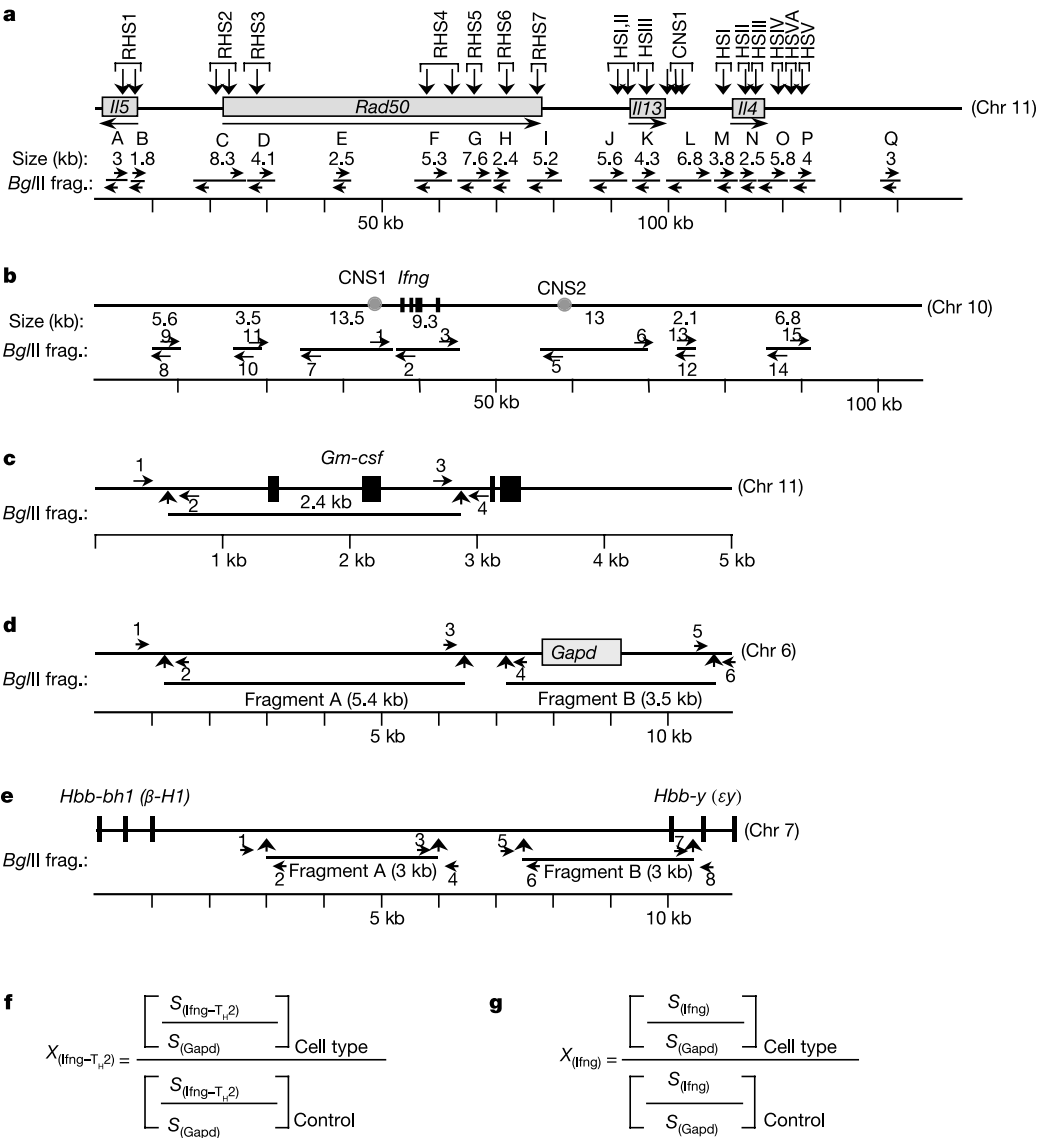


Figure 1 | Spatial organization of the genetic loci used for the 3C analysis. **a**, Schematic representation of the murine T_H2 cytokine locus on mouse chromosome 11. Arrows on the top represent all the DNase I hypersensitive sites characterized so far. Below the locus are presented all the BglII restriction enzyme fragments used in our 3C analysis and their relative position on the locus. **b–e**, Schematic representation of the murine *Ifng* locus (**b**), *Gm-csf* locus (**c**), *Gapd* locus (**d**) and β -globin locus (partial) (**e**). **f**, The equation used for the interchromosomal interactions in the 3C analysis. *X* is the relative crosslinking frequency between two fragments.

g, The equation used for the *Ifng* intrachromosomal interactions in the 3C analysis. For the *Gapd* control the values we obtained by dividing the PCR signal (*S*) detected in a cell tissue with the PCR signal for its control template were 0.45 (naive T cells), 0.63 (T_H1) and 0.72 (T_H2). The values we obtained using the β -globin locus as a control were accordingly 0.46, 0.64 and 0.69, which are comparable to those for the *Gapd* gene. Re-analysing the data using β -globin as a control does not change the interpretation of the results.

C) and *Il5* promoter (fragment B), respectively (Fig. 2a–c). The DNase I hypersensitive site RHS6 of the T_H2 LCR is the only one present in naive T cells, T_H1 and T_H2 cells²⁴; *Rad50* promoter DNase I hypersensitive sites are also constitutively hypersensitive. Notably, in naive T cells the interactions were very strong, with a crosslinking frequency of 4–5, a value similar to that found for the strongest intrachromosomal interactions in the *Il4* locus¹⁸. Upon differentiation of naive T cells to effector T_H1 and T_H2 cells the crosslinking frequencies became much lower. Using the 3C analysis we also determined the interaction of *Ifng* CNS1 and *Ifng* CNS2 fragments with all fragments in the T_H2 locus, but no specific interaction was detected (Fig. 2d). All of the above-mentioned results were highly reproducible over six experiments.

Interchromosomal interactions require RHS7. To determine whether the T_H2 LCR is required for the interchromosomal interactions, a specific region covering DNase I hypersensitive site RHS7 of the LCR was deleted; this region seems to be the most important element of the LCR and is essential for T_H2 cytokine gene expression. Furthermore, in previous studies we found that RHS7 deletion leads to the abrogation of all RHS6 interactions with other *cis*-elements on the *Il4* locus²⁷. Therefore we hypothesized that RHS7 deletion would cause loss of RHS6 interactions with the *Ifng* locus. We prepared templates for the 3C analysis and measured the relative crosslinking frequency between the *Ifng* gene and any fragment in the T_H2 cytokine locus. RHS7 was required for the interchromosomal interaction of the *Ifng* gene fragment with the T_H2 locus. Notably, in naive T cells and to a lesser extent in effector T_H1 and T_H2 cells, the interactions of the *Ifng* gene with the *Il5* promoter, *Rad50* promoter and RHS6 were greatly reduced, from a crosslinking frequency of about 5 to a crosslinking frequency with a value close to 1, the same

value considered to be background in an intrachromosomal interaction (Fig. 2a–c).

To make sure that the results for the interchromosomal interaction of different loci were not an artefact of the 3C technique, we performed experiments analysing the crosslinking frequency between different fragments from several genetic loci in different chromosomes. No interactions were detected by the 3C analysis for restriction fragments A and B of the *Gapd* locus with any fragment in the T_H2 locus. Similarly, interactions were not detected between the seven restriction fragments of the *Ifng* locus and the two restriction fragments in the *Gapd* locus, between the *Ifng* locus and the *Gm-csf* locus, nor between the *Gapd* locus and the *Gm-csf* locus (Fig. 2d).

Inter- to Intrachromosomal interaction switch in effector T cells. Because we found that the interaction between the *Ifng* gene and the specific regions of the T_H2 locus are largely lost during the differentiation of naive T cells to effector T-helper cells, we wanted to see whether new compensating interactions emerge in effector cells that might replace the interchromosomal interactions described. We therefore analysed the intrachromosomal interactions between seven *Bgl*III restriction enzyme fragments in the *Ifng* locus. We found that the restriction enzyme fragment covering the conserved sequence *Ifng* CNS1 upstream of the *Ifng* gene interacted with the *Ifng* gene fragment, as previously described²⁸, but not with the fragment spanning *Ifng* CNS2 (Fig. 3a–c). This interaction was profound in naive T cells, with a crosslinking frequency close to 5; in effector T_H1 and T_H2 cells the interaction was present but was weaker compared to that observed in naive T cells. We next analysed the crosslinking frequencies of the *Ifng* gene with the upstream and downstream conserved regions and the control regions, and we found it to interact

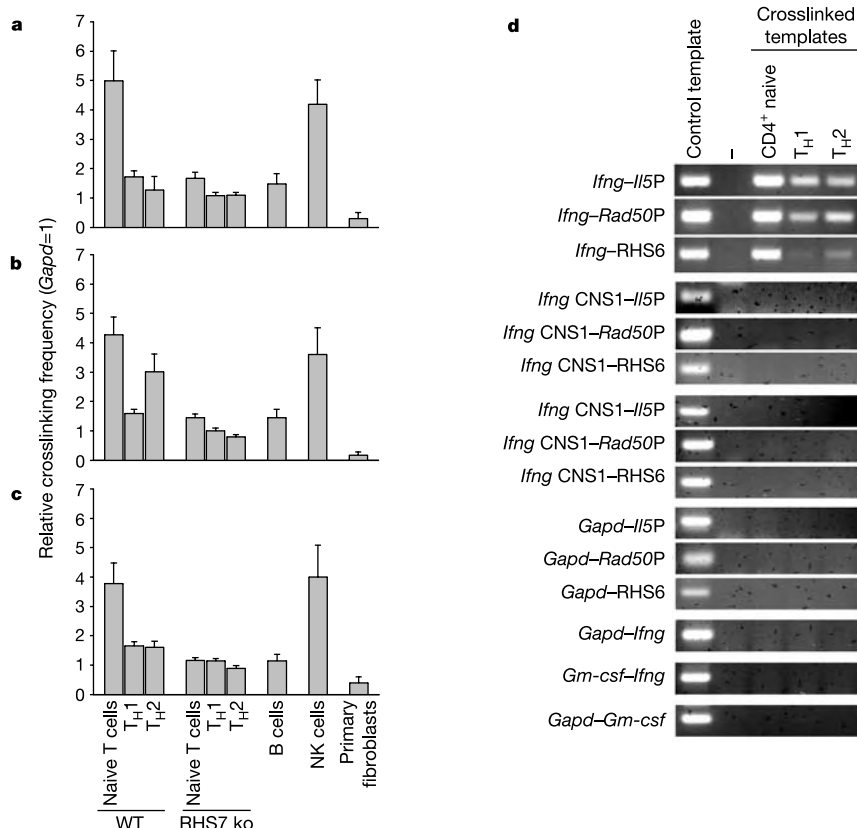


Figure 2 | Interchromosomal, T_H2 LCR-regulated interactions. On the y axis the relative crosslinking frequency between two fragments is presented. The value of 1 is the crosslinking frequency for the two closely linked *Gapd* fragments. **a–c**, Relative crosslinking frequencies for the interchromosomal interaction between the *Ifng* gene and the *Il5* promoter (**a**), the *Rad50*

promoter (**b**) and RHS6 DNase I hypersensitive site of the T_H2 LCR (**c**). Error bars represent the standard deviation between at least three repeats for each of two cell preparations. **d**, PCR signals detected from the 3C analysis for naive T cells, T_H1 and T_H2 cells using pairs of primers for the regions indicated on the left.

with *Ifng* CNS1 in all three cell types (Fig. 3d–f), and to specifically interact with *Ifng* CNS2 in T_H1 cells (Fig. 3e). The *Ifng* CNS2 fragment was associated with the *Ifng* gene only in T_H1 cells (Fig. 3h). Although the crosslinking frequency value of 2.1 ± 0.35 detected for the latter interaction is only double that of the value 1, set for the background (based upon *Gapd*–*Gapd* interactions), we consider this significant because this association was only detected in T_H1 cells and not in naive T cells nor T_H2 cells, where the interactions were in fact completely undetectable (that is, much lower even than the *Gapd* background). There were no interactions detected between IFN- γ CNS1 and IFN- γ CNS2.

Co-localization of the *Ifng* and T_H2 loci revealed by FISH. To confirm the interchromosomal interactions between the *Ifng* and T_H2 loci, fluorescence *in situ* hybridization (FISH) experiments were performed. Hybridizations were performed using sorted naive $CD4^+$ T cells or T_H1 and T_H2 cells, differentiated for 5 days, from wild-type or $RHS7^{-/-}$ mice. First, we performed the experiments for wild-type naive T cells, T_H1 and T_H2 cells using two-dimensional FISH. These experiments were repeated twice using two independent preparations of wild-type T cells (Supplementary Fig. 1). These FISH data confirmed our results obtained using the 3C technique. In about 50% of wild-type naive T cells one allele of the T_H2 locus was

co-localized with one allele of the *Ifng* locus, whereas this interaction was considerably decreased in effector T_H1 and T_H2 cells. As an important control, we examined the co-localization of the *Gapd* locus (located in chromosome 6) with the T_H2 locus or the *Ifng* locus. In all cell types the percentage of cells having one allele for the T_H2 or *Ifng* gene loci co-localized with the *Gapd* locus was always low and is probably the background level of the technique, because it is not confirmed with the 3C method (Supplementary Fig. 1). We wanted to expand further our analysis after these findings; therefore we used a FISH technique where the three dimensional structure of the cells is maintained. We repeated the experiments for the wild-type and $RHS7^{-/-}$ cells (Fig. 4a–f). In three independent experiments with two preparations of cells we observed that the percentage of naive $CD4^+$ T cells having at least one allele of the T_H2 locus co-localized with one allele of the *Ifng* locus was close to 40%, compared with only 10–13% observed in effector T_H1 and T_H2 cells (Fig. 4a, c, e). In $RHS7^{-/-}$ naive T cells the percentage of cells with at least one allele co-localized was increased to 71.5%. However, more careful analysis of the cells revealed that in wild-type cells the signals scored as co-localized had a distance of ≤ 5 pixels, but in more than 85% of $RHS7^{-/-}$ naive T cells that showed co-localization, the distance between the co-localized signals was greater (about 5–8 pixels),

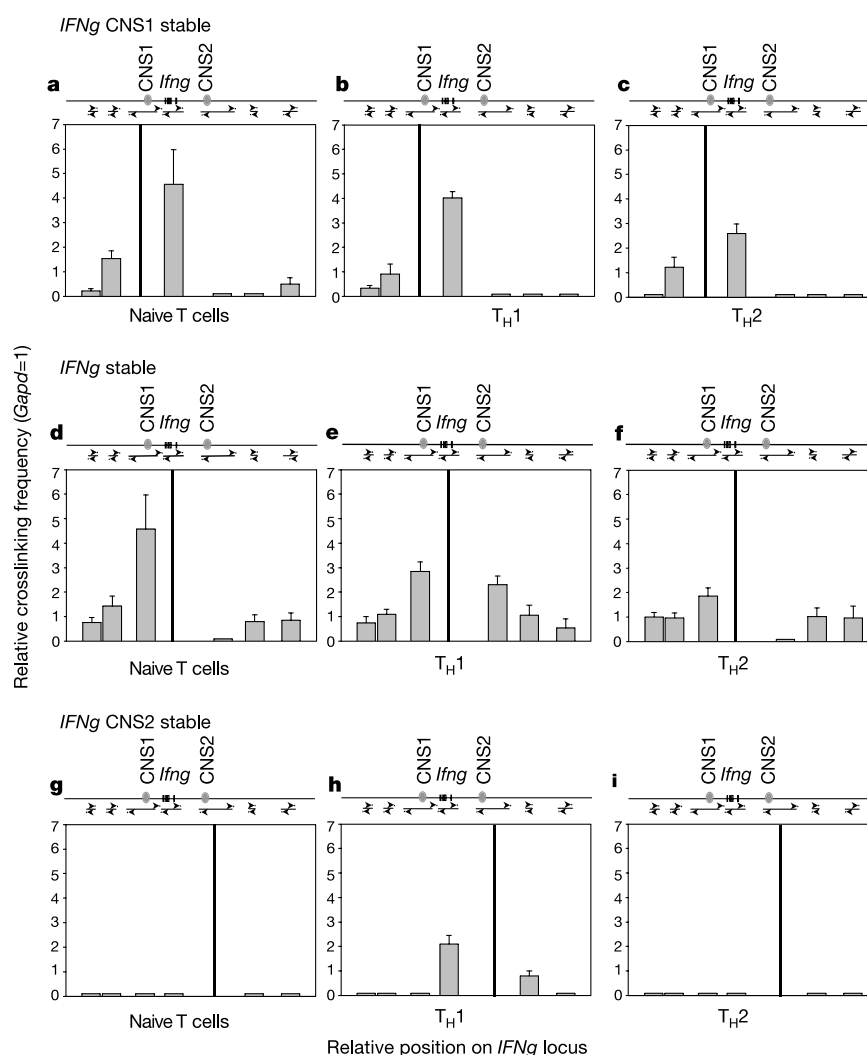


Figure 3 | Intrachromosomal enhancer-promoter interactions in the *Ifng* gene. a–i, 3C analysis for the *Ifng* locus (a–c), CNS1 (d–f) and CNS2 (g–i). Error bars represent the standard deviation between four repeats of all PCR reactions. The x axis represents the relative position of fragments on the *Ifng* locus; the y axis represents the relative crosslinking frequency between two

fragments. The value of 1 is the crosslinking frequency for the two closely located GAPD restriction enzyme fragments. ‘Stable’ means that the pair of primers for the designated restriction fragment was combined with the pair of primers of the other fragments.

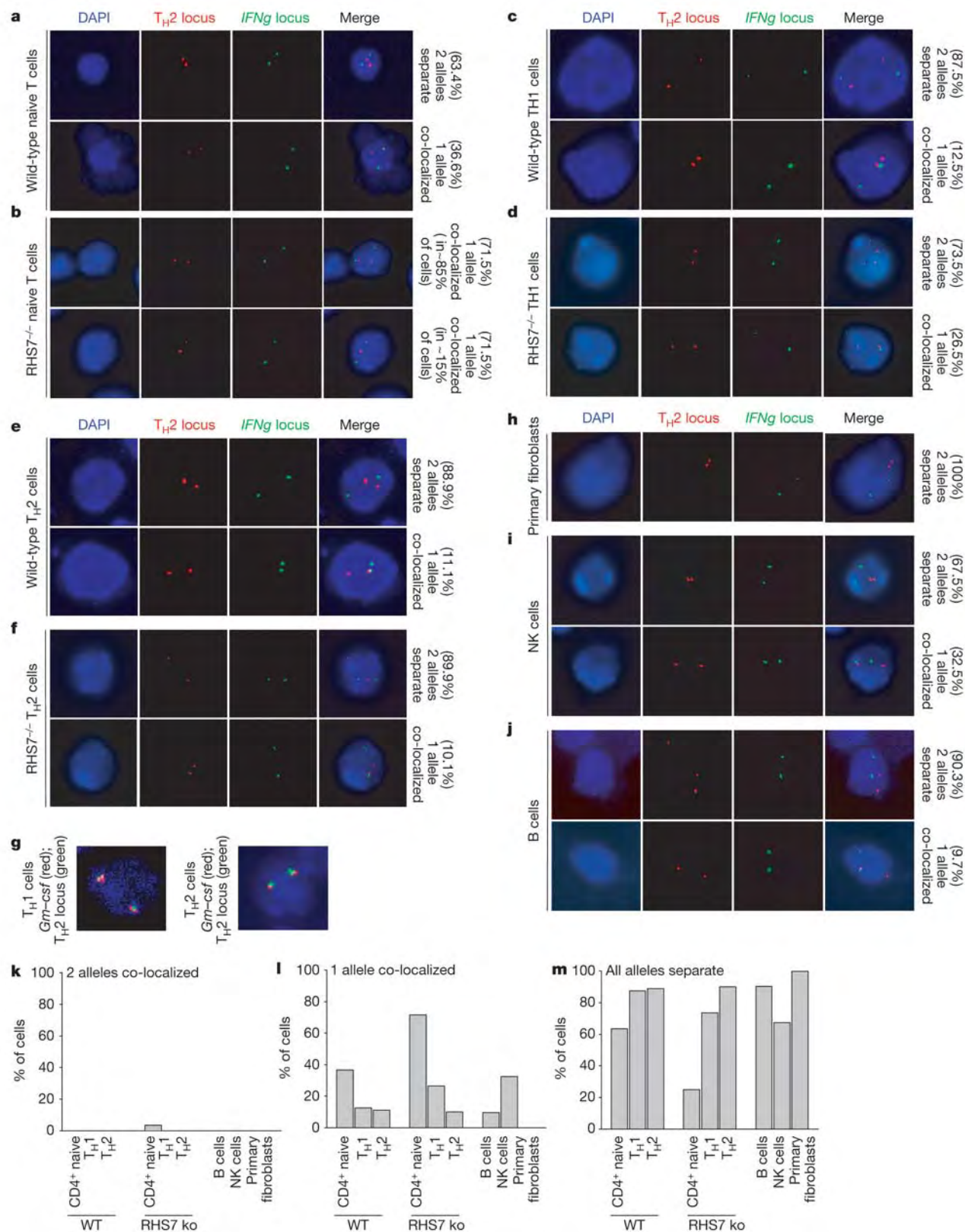


Figure 4 | Co-localization of the *Ifng* and T_H2 loci as revealed by FISH. Hybridization of naive T cells (**a**, **b**), T_H1 cells (**c**, **d**), T_H2 cells (**e**, **f**), primary fibroblasts (**h**), NK cells (**i**) and B cells (**j**), with rhodamine-dUTP-labelled probe for the T_H2 locus and SpectrumGreen-dUTP-labelled probe for the *Ifng* locus. The first column of photomicrographs ($\times 100$) represents DAPI staining of the nuclei for the presence of DNA. **g**, Hybridization of T_H1 and T_H2 cells with probes for *Gm-csf* and T_H2 loci. **k–m**, Diagrammatic representation of the percentage of cells with co-localization. The following number of cells were scored for each cell type: wild-type naive T cells, 235; wild-type T_H1 , 149; wild-type T_H2 , 125; $RHS7^{-/-}$ naive, 133; $RHS7^{-/-}$ T_H1 ,

84; $RHS7^{-/-}$ T_H2 , 93; primary fibroblasts, 52; B cells, 54; NK cells, 73. Signals with a relative distance of ≤ 5 pixels were scored as co-localized for all cell types with the exception of $RHS7^{-/-}$ cells, where the distance of co-localized signals was ≤ 8 pixels. The average diameter (in pixels) of each cell nucleus was: (presented as the mean $\pm$ s.d.) naive cells 126.58 ± 34.03 ; T_H1 172.46 ± 21.75 ; T_H2 146.06 ± 31.67 ; B cells 139.1 ± 18.67 ; NK cells 135.05 ± 19.68 ; primary fibroblasts 184.23 ± 24.7 . All cells for each category were scored blindly and slides were prepared from two different preparations of cells.

implying that the *Ifng* and T_H2 loci still co-localize in $RHS7^{-/-}$ cells, but in a looser manner. As a positive control for co-localization we hybridized cells with probes for the T_H2 locus and the *Gm-csf* gene locus, both of which are on the same chromosome about 600 kb apart. In this case all signals were co-localized (Fig. 4g). We also quantified the distance between the co-localized *Ifng* and *Il4* gene regions in an additional random sample of cells. Although these two loci were 1.87 ± 1.97 pixels apart in wild-type naive T cells, the distance (4.45 ± 2.37) was two–three times greater in $RHS7^{-/-}$

naive T cells. This difference was highly significant ($P < 0.001$). In total, we analysed approximately 1,000 cells of wild-type and $RHS7$ knockout mice in eight experiments. Thus, by two independent approaches (3C and FISH), interchromosomal interactions were detected between the *Ifng* and T_H2 loci in naive T cells, and were reduced in effector T cells.

We also analysed the interaction of the *Ifng* and the T_H2 loci for natural killer (NK) cells, B cells and primary fibroblasts (Fig. 4h–j). NK cells, which have the potential to express IFN- γ , also show a

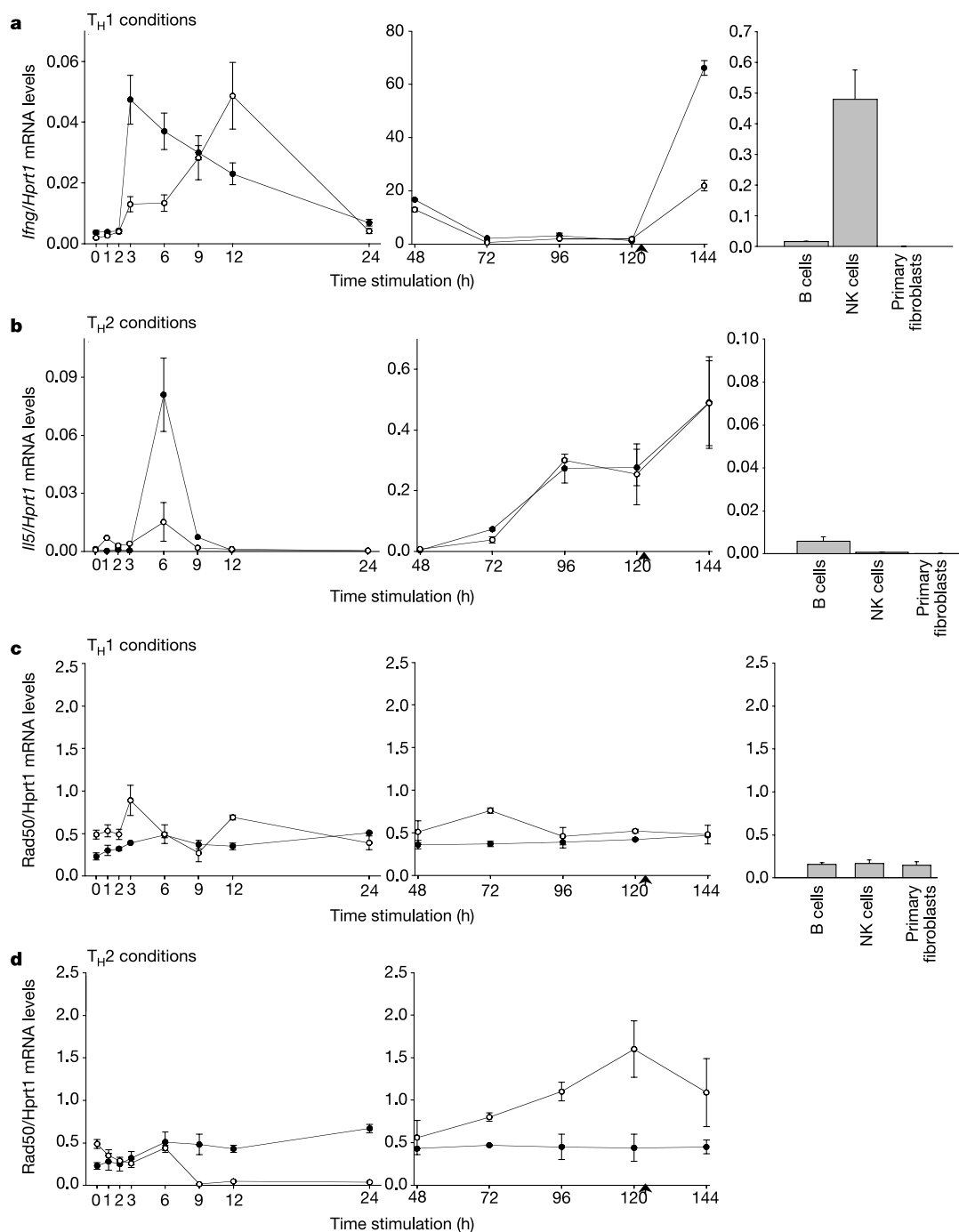


Figure 5 | Deletion of RHS7 hypersensitive site on chromosome 11 affects the expression of *Ifng* on chromosome 10. **a**, *Ifng* mRNA expression levels corrected for the expression levels of *Hprt1* (y axis) for wild-type (filled circles) and $RHS7^{-/-}$ (open circles) sorted $CD4^+$ naive T cells differentiated under T_H1 conditions. The corrected *Ifng* mRNA expression levels of B cells, NK cells and primary fibroblasts are indicated on the right.

b, *Il5* mRNA expression levels of cells differentiated under T_H2 conditions. **c**, **d**, *Rad50* mRNA expression levels of cells differentiated under T_H1 (**c**) or T_H2 (**d**) conditions. Error bars represent the standard deviation of two repeats each of three preparations of cells, with real time RT-PCR reactions performed in duplicates.

high degree of association between the *Ifng* and T_H2 loci in 32.5% of the cells scored (Fig. 4l). Non-expressing B cells and fibroblasts, however, showed a low degree of association. A summary of the percentages of cells with co-localized or separate signals is depicted in Fig. 4k–m.

We examined whether the localization of the T_H2 and *Ifng* loci, and especially the co-localized loci, were located in an active or an inactive nuclear compartment. We compared co-localization of the cytokine loci with DAPI (4,6-diamidino-2-phenylindole) staining for DNA (which highlights condensed heterochromatic regions), and performed immuno-DNA FISH experiments using an antibody for heterochromatin protein 1 alpha (HP1- α), a protein that is mainly located in heterochromatic regions^{29,30}. Only 7.1% of the co-localized loci in wild-type naive T cells and 8.8% of the co-localized loci in *RHS7*^{-/-} naive T cells were located in heterochromatic regions (Supplementary Fig. 2). On the basis of this finding we propose that co-localization of the T_H2 and the *Ifng* loci poises or prepares these loci for their rapid expression upon stimulation^{31–33} rather than holding them in a repressive environment.

Interchromosomally interacting loci are poised for transcription. Upon *in vitro* stimulation of $CD4^+$ naive T cells under T_H1 and/or T_H2 conditions there is an early production of cytokines, with the peak of their expression at 2–3 h after stimulation, which occurs days before the naive cells differentiate to T_H1 or T_H2 cells⁵. We sorted $CD4^+$ naive T cells from wild-type and *RHS7*^{-/-} mice and stimulated them *in vitro* with plate-bound anti-CD3 and anti-CD28 antibodies under T_H1 or T_H2 conditions. At the indicated time points (Fig. 5a–d) cells were collected, and upon preparation of RNA and complementary DNA, real-time RT-PCR analysis was performed. Upon stimulation of the cells there was a rapid, transient peak of transcription at 3 h for *Ifng* production under T_H1 conditions for the wild-type cells, whereas this peak was delayed by 9 h for the *RHS7*^{-/-} cells. After 5 days of stimulation, by which time we obtained differentiated effector T_H1 and T_H2 cells, we re-stimulated the T_H1 cells for 24 h; the production of *Ifng* was reduced threefold in the *RHS7*^{-/-} T cells (Fig. 5a). Under the same stimulation conditions the early peak of transcription for the *Il5* gene was 6 h after stimulation under T_H2 conditions for wild-type cells, and was decreased 5.3-fold for the *RHS7*^{-/-} cells under the same stimulation conditions. There was no difference in production of *Il5* at late time points after stimulation for the wild-type and *RHS7*^{-/-} cells (Fig. 5b). *Rad50* mRNA expression did not reveal any significant difference in the expression of the gene in the early compared to the late time points after stimulation (Fig. 5c,d). Thus, by several criteria, deletion of a hypersensitive site in the T_H2 locus affects transcription of the *Ifng* locus located on a different chromosome but connected to the T_H2 locus through an interchromosomal complex.

Discussion

Our study shows that there is a direct interchromosomal interaction between two loci that are expressed as mutually exclusive alternatives in two different cell types. T_H1 cells express IFN- γ on chromosome 10 and T_H2 cells express IL-4, IL-5 and IL-13—the three interleukins residing as linked genes in the T_H2 cytokine locus—on chromosome 11; these three interleukins are coordinately regulated by the action of the T_H2 LCR and other *cis* elements. The co-localization of the *Ifng* locus with the T_H2 locus was established with the 3C technique and confirmed with FISH experiments. The 3C interchromosomal interactions were extremely strong: they were as strong as the tightest intrachromosomal interactions that we previously described for the T_H2 locus¹⁸.

The functional significance of the interchromosomal association is of great interest. We think that it is significant that naive $CD4^+$ T cells have the potential to rapidly express both T_H1 and T_H2 cytokines at low levels immediately after activation⁵, and we hypothesize that this is a consequence of the ‘poised chromatin hub’ configuration formed between the *Ifng* and T_H2 cytokine loci, which robustly interact with

each other. The T_H2 LCR, and more specifically the protein modules recruited to the hypersensitive sites comprising the LCR, are of great importance for this phenomenon, because the deletion of the *RHS7* site of the LCR significantly delays (9 h for *Ifng* expression) or reduces (5.3 times for the *Il5* expression) the early expression of the cytokines involved in the interchromosomal interactions. The interchromosomally interacting loci are located in a transcriptionally positive environment, and may recruit remodelling complexes or acetyltransferases to form a positive environment for the early expression of cytokines³⁴. The fact that many nuclei of naive T cells have one co-localized pair of alleles may relate to the phenomenon of mono-allelic expression, whereby cytokines and other genes are expressed from only one of the two alleles^{32,35}. The concept that expression might initiate first at the co-localized allele is attractive. Future RNA-FISH experiments in conjunction with DNA-FISH will help us to identify which allele is expressed first and hence determine the inter-relationships of these phenomena.

We believe that there is a dynamic interplay between partners, which is reflected by the interactions we describe. Upon differentiation of naive T cells to effector cells the interactions change: in T_H1 cells the interchromosomal association between the *Ifng* gene and the T_H2 cytokine locus is significantly reduced and is apparently replaced by another T_H1 -specific interaction between the *Ifng* gene and *Ifng* CNS2, a tissue-specific transcriptional enhancer. However, we cannot be sure that this is a true ‘exchange’ of interactions, because with the 3C technique we detect interactions between genomic loci in a population of cells at a given time point. The structure that we describe suggests that elements located in one chromosome may positively or negatively regulate the expression of genes in a locus in another chromosome. The action of LCRs may therefore impact on not only the activation of adjacent genes but also the activation or repression of genes in other loci in the genome. Put another way, the T_H2 LCR seems to regulate not only the transcription of the T_H2 cytokines but also *Ifng* gene expression. On the other hand, the T_H2 LCR may control the developmentally regulated positioning of the T_H2 locus to the so-called transcriptional factories or other compartments of the nucleus that facilitate its expression.

The interchromosomal associations described in this work are unlikely to be restricted to T-helper cell differentiation but rather, may be a general phenomenon occurring at multiple genetic loci in the cell nucleus, particularly in gene systems where coordinate or alternative regulation is important. Potential examples include olfactory receptor genes, heavy and light chain genes of immunoglobulins, and α - and β -globin genes, where gene expression or rearrangement might be coordinated and rendered monoallelic. Furthermore, we believe that such interactions may be important in disease. It has been proposed that position in the nucleus may control gene expression³⁶. Proximity of two chromosomes in the same general region of the nucleus has also been suggested to contribute to rearrangements, which lead to malignancy³⁷. For example, chronic myelogenous leukaemia (CML) is marked by the presence of a distinct cytogenetic abnormality that results from a translocation between chromosomes 9 and 22, known as the Philadelphia chromosome. Also, in murine plasmacytomas (MPCs) dysregulation of the *Myc* transcript is achieved by chromosomal translocation that juxtaposes the *c-Myc/Pvt-1* locus on chromosome 15 with one of the immunoglobulin loci on chromosomes 12 (IgH), 6 (Ig κ) or 16 (Ig λ). We propose that the developmentally regulated co-localization of multiple loci such as *Bcr* and *Abl* in the cell nucleus might result in these malignancies. Future analysis of interchromosomal interactions between candidate loci may shed light on the higher-order regulation of expression in a cell nucleus.

METHODS

Mice and cell cultures. C57BL/6 mice were derived from Jackson Laboratories. Ten to twenty spleens from 4–6-week-old C57BL/6 mice were used to make

single-cell suspensions. Isolation and differentiation of naive CD4⁺ T cells was done as described previously³⁸. Naive cells (1×10^6 per ml) were stimulated with plate-bound anti-CD3 (145-2C11 clone, American Type Culture Collection) and anti-CD28 mouse antibodies (Pharmingen). B cells and NK cells were isolated as described before¹⁸. Experiments performed in this study were approved by the Yale University Institutional Animal Care and Use Committee.

3C analysis. The protocol that was used for the 3C analysis is described elsewhere¹⁸ (see also Supplementary Methods). The percentage of digestion in each different cell type was identified using real-time PCR analysis with primers spanning the *Bgl*II sites of interest (Supplementary Fig. 3). To generate control templates for the positive controls, bacterial artificial chromosome (BAC) clones were used for all the loci of interest. For the *T_H2* locus we used the BAC clone B182 (Genome systems), for *Ifng* locus we used the BAC clone RPCL24-352N22 (CHORI), for *Gm-csf* we used the BAC clone RPCL23-397C23 (CHORI), and for *Gapd* and β -globin loci we used PCR to amplify the regions spanning the *Bgl*II sites of interest, as described previously¹⁸. For the templates to function as positive controls for the interchromosomal 3C analysis, different BAC clones were mixed in equimolar quantities, digested and ligated. The templates were prepared two times from independent experiments and the whole set of PCR reactions was repeated at least three times. Data presented are the average of results for all PCR reactions done. An important experimental control was introduced in our analysis. All PCR signals that were considered as positive were isolated, gel extracted (Qiaquick, Qiagen), cloned in a TA vector (TA cloning kit, Invitrogen) and sequenced to identify unambiguously the sequence of the chimaeric *Bgl*II ligated fragments. The primers for the 3C analysis are available upon request.

Two-dimensional FISH. Sorted naive T cells, *T_H1* and *T_H2* cells were fixed in methanol:acetic acid (3:1), spotted on glass slides and air dried. Two micrograms of BAC DNA (the same BAC clones as the 3C analysis and for *Gapd* locus: RP23-216N11, CHORI) was labelled using the nick translation kit (Roche), according to the manufacturer's instructions, and either 0.05 mM rhodamine-dUTP (Roche) or 0.025 mM SpectrumGreen-dUTP (Vysis). The probe was incubated overnight (15 h) at 15 °C and then purified using the PCR purification kit (Qiagen) and eluted in 30 μ l of water. The probe mix consisted of 7 μ l master mix (60% formamide, 2 $\times$ SSC, 30% dextran sulphate), 1 μ l mouse cot-1 DNA (Roche) and 2 μ l of each probe. The signals were detected using an Olympus BX51 microscope equipped with Vysis filters. The data presented in Supplementary Fig. 1 have been generated using this protocol.

Three-dimensional FISH. Cells were prepared on coverslips with the following protocol, which permits the maintenance of the three-dimensional structure of the cells. Cells were permitted to attach onto poly-L-lysine-coated coverslips (BD BioCoat Cellware poly-L-lysine, 12 mm; BD Biosciences) and were fixed with 4% PFA in 0.3 $\times$ PBS for 12 min, washed three times with 1 $\times$ PBS (5 min per wash) and permeabilized with 0.5% Triton X-100 in PBS for 10 min. After an incubation of 30 min in 20% glycerol/PBS, cells were freeze-thawed three times in liquid nitrogen. Cells were incubated in 0.1N HCl for 5 min and rinsed in 2 $\times$ SSC. After the fixation and pre-treatment, the cells were stored in 50% formamide/2 $\times$ SSC. The same probes that were used for the two-dimensional FISH were also used for three-dimensional FISH in an overnight incubation. After washing the coverslips we proceeded with the immuno-FISH. An anti-HP1- α monoclonal antibody (MAB3584, Chemicon International) was used in a dilution of 1/400. The secondary capture antibody (A-11068, Alexa Fluor 350 goat anti-mouse, Molecular Probes) was used in a dilution of 1/200. Coverslips were mounted in VectaShield (Vector Laboratories) and were visualized using an automated Olympus BX61 microscope. All the data presented in Figs 1–5 as well as Supplementary Fig. 2 have been generated with the protocol described here.

Real-time PCR analysis. Quantitative reverse transcriptase (RT)-PCR was performed using real-time fluorogenic 5'-nuclease PCR using an i-cycler (Biorad) according to the manufacturer's instructions. The primers and probes used for the amplification of *Ifng*, *Il5* and *Hprt1* are described elsewhere³. The primers and probe for the amplification of *Rad50* cDNA was Mm00485504_m1 (Applied Biosystems). Cytokine transcripts were normalized to *Hprt1* abundance. For the analysis of the percentage of the restriction enzyme digestion of templates with *Bgl*II (Supplementary Fig. 3 online) we used iQ Sybr Green Supermix (Biorad) and pairs of primers spanning the *Bgl*II sites and amplifying products of a size of 80–160 base pairs (bp). The sequences of primers are available upon request.

Received 11 February; accepted 30 March 2005.

Published online 8 May 2005.

- Cremer, T. & Cremer, C. Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nature Rev. Genet.* **2**, 292–301 (2001).

- Spector, D. L. The dynamics of chromosome organization and gene regulation. *Annu. Rev. Biochem.* **72**, 573–608 (2003).
- Brown, K. E. *et al.* Association of transcriptionally silent genes with Ikaros complexes at centromeric heterochromatin. *Cell* **91**, 845–854 (1997).
- Brown, K. E., Baxter, J., Graf, D., Merkenschlager, M. & Fisher, A. G. Dynamic repositioning of genes in the nucleus of lymphocytes preparing for cell division. *Mol. Cell* **3**, 207–217 (1999).
- Grogan, J. L. *et al.* Early transcription and silencing of cytokine genes underlie polarization of T helper cell subsets. *Immunity* **14**, 205–215 (2001).
- Ragoczy, T., Telling, A., Sawado, T., Groudine, M. & Koxak, S. T. A genetic analysis of chromosome territory looping: diverse roles for distal regulatory elements. *Chromosome Res.* **11**, 513–525 (2003).
- Francastel, C., Walters, M. C., Groudine, M. & Martin, D. I. A functional enhancer suppresses silencing of a transgene and prevents its localization close to centromeric heterochromatin. *Cell* **99**, 259–269 (1999).
- Schubeler, D. *et al.* Nuclear localization and histone acetylation: a pathway for chromatin opening and transcriptional activation of the human β -globin locus. *Genes Dev.* **14**, 940–950 (2000).
- Gerasimova, T. I., Byrd, K. & Corces, V. A chromatin insulator determines the nuclear localization of DNA. *Mol. Cell* **6**, 1025–1035 (2000).
- Cobb, B. S. *et al.* Targeting Ikaros to pericentromeric heterochromatin by direct DNA binding. *Genes Dev.* **14**, 2146–2160 (2000).
- Carmo-Fonseca, M., Platani, M. & Swedlow, J. R. Macromolecular mobility inside the cell nucleus. *Trends Cell Biol.* **12**, 491–495 (2002).
- Carmo-Fonseca, M. The contribution of nuclear compartmentalization to gene regulation. *Cell* **108**, 513–521 (2002).
- Jenuwein, T. & Allis, C. D. Translating the histone code. *Science* **293**, 1074–1080 (2001).
- Mahy, N. L., Perry, P. E., Gailchrist, S., Baldock, R. A. & Bickmore, W. A. Spatial organization of active and inactive genes and noncoding DNA within chromosome territories. *J. Cell Biol.* **157**, 579–589 (2004).
- Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing chromosome conformation. *Science* **295**, 1306–1311 (2002).
- Tolhuis, B., Palstra, R.-J., Splinter, E., Grosveld, F. & de Laat, W. Looping and interaction between hypersensitive sites in the active β -globin locus. *Mol. Cell* **10**, 1453–1465 (2002).
- Palstra, R.-J. *et al.* The β -globin nuclear compartment in development and erythroid differentiation. *Nature Genet.* **35**, 190–194 (2003).
- Spilianakis, C. & Flavell, R. A. Long-range intrachromosomal interactions in the T helper type 2 cytokine locus. *Nature Immunol.* **5**, 1017–1027 (2004).
- Takemoto, N. *et al.* TH2-specific DNase I-hypersensitive sites in the murine IL-13 and IL-4 intergenic region. *Int. Immunol.* **10**, 1981–1985 (1998).
- Agarwal, S. & Rao, A. Modulation of chromatin structure regulates cytokine gene expression during T cell differentiation. *Immunity* **9**, 765–775 (1998).
- Loots, G. G. *et al.* Identification of a coordinate regulator of interleukins-4, 13 and 5 by cross-species sequence comparisons. *Science* **288**, 136–140 (2000).
- Agarwal, S., Avni, O. & Rao, A. Cell-type-restricted binding of the transcription factor NFAT to a distal IL-4 enhancer *in vivo*. *Immunity* **12**, 643–652 (2000).
- Lee, G. R., Fields, P. E., Griffin, T. J. IV & Flavell, R. A. Regulation of the Th2 cytokine locus by a Locus Control Region. *Immunity* **19**, 145–153 (2003).
- Fields, P. E., Lee, G. R., Kim, S. T., Bartsevich, V. & Flavell, R. A. Th2-specific chromatin remodeling and enhancer activity in the Th2 cytokine locus control region. *Immunity* **21**, 865–876 (2004).
- Lee, D. U., Avni, O., Chen, L. & Rao, A. A distal enhancer in the IFN- γ locus revealed by genome sequence comparison. *J. Biol. Chem.* **279**, 4802–4810 (2004).
- Shnyreva, M. *et al.* Evolutionarily conserved sequence elements that positively regulate IFN- γ expression in T cells. *Proc. Natl Acad. Sci. USA* **101**, 12622–12627 (2004).
- Lee, G. R., Spilianakis, C. & Flavell, R. A. Hypersensitive site 7 of the Th2 locus control region is essential for expressing Th2 cytokine genes and for long-range intrachromosomal interactions. *Nature Immunol.* **6**, 42–48 (2005).
- Eivazova, E. R. & Aune, T. M. Dynamic alterations in the conformation of the *Ifng* gene region during T helper cell differentiation. *Proc. Natl Acad. Sci. USA* **101**, 251–256 (2004).
- Taddei, A., Maison, C., Roche, D. & Almouzni, G. Reversible disruption of pericentric heterochromatin and centromere function by inhibiting deacetylases. *Nature Cell Biol.* **3**, 114–120 (2001).
- Schmiedeberg, L., Weisshart, K., Diekmann, S., Meyer zu Horst, G. & Hemmerich, P. High- and low-mobility populations of HP1 in heterochromatin of mammalian cells. *Mol. Biol. Cell* **15**, 2819–2833 (2004).
- Eggenschwiler, J. *et al.* Mouse mutant embryos overexpressing IGF-II exhibit phenotypic features of the Beckwith–Wiedemann and Simpson–Golabi–Behmel syndromes. *Genes Dev.* **11**, 3128–3142 (1997).
- Dong, C. & Flavell, R. A. Cell fate decision: T-helper 1 and 2 subsets in immune responses. *Arthritis Res.* **2**, 179–188 (2000).
- Guo, L. *et al.* In Th2 cells the *Il4* gene has a series of accessibility states associated with distinctive probabilities of IL-4 production. *Proc. Natl Acad. Sci. USA* **99**, 10623–10628 (2002).
- Kosak, S. T. & Groudine, M. Form follows function: the genomic organization of

- cellular differentiation. *Genes Dev.* **18**, 1371–1384 (2004).
35. Guo, L. *et al.* In TH2 cells the IL-4 gene has a series of accessibility states associated with distinctive probabilities of IL-4 production. *Proc. Natl Acad. Sci. USA* **99**, 10623–10628 (2002).
36. Parada, L. A., Sotiriou, S. & Misteli, T. Spatial genome organization. *Exp. Cell Res.* **296**, 64–70 (2004).
37. Roix, J. J., McQueen, P. G., Munson, P. J., Parada, L. A. & Misteli, T. Spatial proximity of translocation-prone gene loci in human lymphomas. *Nature Genet.* **34**, 287–291 (2003).
38. Avni, O. *et al.* TH cell differentiation is accompanied by dynamic changes in histone acetylation of cytokine genes. *Nature Immunol.* **3**, 643–651 (2002).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank Wyeth Laboratories for donation of IL-12. We

also thank F. Manzo for assistance with manuscript preparation, T. Gerasimova for suggestions with the FISH protocols, and D. Sakkas for use of his fluorescence microscope. We are also grateful to F. G. Grosveld and W. de Laat for originally providing us with detailed protocols and help with establishing the 3C technique. We would like to thank C. Szekely for assistance with graphs. C.S. is supported by a Cancer Research Institute fellowship; M.D.L. is partly supported by a Human Frontiers Science Program long-term fellowship; T.T. is supported by a Ruth L. Kirchstein NIH/NRSA/NIA post-doctoral fellowship; R.A.F. is an Investigator of the Howard Hughes Medical Institute.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to R.A.F. (richard.flavell@yale.edu).

ARTICLES

Highly efficient endogenous human gene correction using designed zinc-finger nucleases

Fyodor D. Urnov¹, Jeffrey C. Miller¹, Ya-Li Lee¹, Christian M. Beausejour¹, Jeremy M. Rock¹, Sheldon Augustus¹, Andrew C. Jamieson¹, Matthew H. Porteus², Philip D. Gregory¹ & Michael C. Holmes¹

Permanent modification of the human genome *in vivo* is impractical owing to the low frequency of homologous recombination in human cells, a fact that hampers biomedical research and progress towards safe and effective gene therapy. Here we report a general solution using two fundamental biological processes: DNA recognition by C₂H₂ zinc-finger proteins and homology-directed repair of DNA double-strand breaks. Zinc-finger proteins engineered to recognize a unique chromosomal site can be fused to a nuclease domain, and a double-strand break induced by the resulting zinc-finger nuclease can create specific sequence alterations by stimulating homologous recombination between the chromosome and an extrachromosomal DNA donor. We show that zinc-finger nucleases designed against an X-linked severe combined immune deficiency (SCID) mutation in the *IL2R γ* gene yielded more than 18% gene-modified human cells without selection. Remarkably, about 7% of the cells acquired the desired genetic modification on both X chromosomes, with cell genotype accurately reflected at the messenger RNA and protein levels. We observe comparably high frequencies in human T cells, raising the possibility of strategies based on zinc-finger nucleases for the treatment of disease.

Most human monogenic disorders remain difficult to treat because therapeutic transgenes do not undergo homologous recombination (HR) into the mutated locus^{1,2}, and gene addition by virus-driven random integration remains a challenge owing to transgene silencing, improper activity or misintegration^{3,4}. Furthermore, targeted alteration of DNA sequence *in vivo*—in principle, a powerful basic research technique for studying genome function—in mammals requires sophisticated targeting vectors and drug-based selection^{1,2}, which limits the use of this approach^{5–7}.

The C₂H₂ zinc-finger, originally discovered in *Xenopus*⁸, is the most common DNA binding motif in all metazoa⁹. Each finger recognizes 3–4 base pairs of DNA via a single α -helix^{10,11}, and several fingers can be linked in tandem to recognize a broad spectrum of DNA sequences with high specificity^{12–14}. Engineered zinc-finger protein (ZFP)-based DNA binding domains with novel specificities have been extensively applied *in vivo* to target various effector domains^{12,15}. Work from the Chandrasegaran laboratory has shown that a ZFP can be coupled to the nonspecific DNA cleavage domain of the Type IIS restriction enzyme, *FokI*, to produce a zinc-finger nuclease (ZFN)¹⁶, which then cuts the DNA sequence determined by the ZFP^{16,17}. An important specificity mechanism derives from the requirement that two ZFNs bind the same locus, in a precise orientation and spacing relative to each other, to create a double-strand break (DSB; Fig. 1a)¹⁷. One mechanism by which eukaryotic cells heal DSBs is homology-directed repair (Fig. 1b)^{18–20}, which transfers information missing at the break from a homologous DNA molecule (Fig. 1b). Work from the Jasin laboratory²¹, followed by that of others^{22,23}, demonstrated that the endonuclease I-SceI can potentiate HR into loci previously engineered to contain its own recognition site, and the Carroll^{24,25} and Baltimore²⁶ laboratories have

shown that a ZFN-invoked DSB increases the rate of HR in model systems.

Here we invoke this process at an endogenous locus in the human genome. We show that a DSB induced by engineered ZFNs at an inherited disease mutation hotspot rapidly leads to permanent, precise modification of up to 20% of the chromatids. Such unprecedented HR frequency, combined with the ability to engineer ZFNs against essentially any sequence, establish the usefulness of ZFN-driven genome editing as a tool for human somatic cell genetics, and also illuminate the potential for gene correction therapy of human inherited disease.

Optimization of ZFN-driven gene correction

To rapidly gauge the potential for ZFN-driven alteration of the genome *in vivo*, we elaborated on published work²⁶ and established a reporter system for gene correction (Fig. 1c). We used an archive of engineered zinc-finger motifs¹⁵ to assemble two ZFNs against the gene encoding enhanced green fluorescent protein (eGFP). The DNA stretch between the ZFP binding sites was then disabled by a stop codon and a frameshift, and the resulting nonfunctional transgene was stably integrated at a single location (data not shown) in the genome of HEK 293 cells. When a fragment of wild-type eGFP corresponding to the mutation-disabled stretch was introduced into this reporter cell line as plasmid DNA, approximately 1–2 cells in 500,000 restored GFP function (left column of Fig. 1c: donor only), as expected on the basis of published measurements². In contrast, four days after the simultaneous introduction of the donor and ZFN expression constructs, 2.2% of asynchronously growing cells and 10.2% of cells arrested for 30 hours at the G2/M cell cycle boundary were GFP-positive (right column of Fig. 1c: donor + ZFNs). These

¹Sangamo BioSciences, Inc. Pt. Richmond Tech Center 501, Canal Blvd, Suite A100 Richmond, California 94804, USA. ²Department of Pediatrics and Department of Biochemistry, University of Texas Southwestern Medical Center, 5323 Harry Hines Blvd, Dallas, Texas 75390, USA.

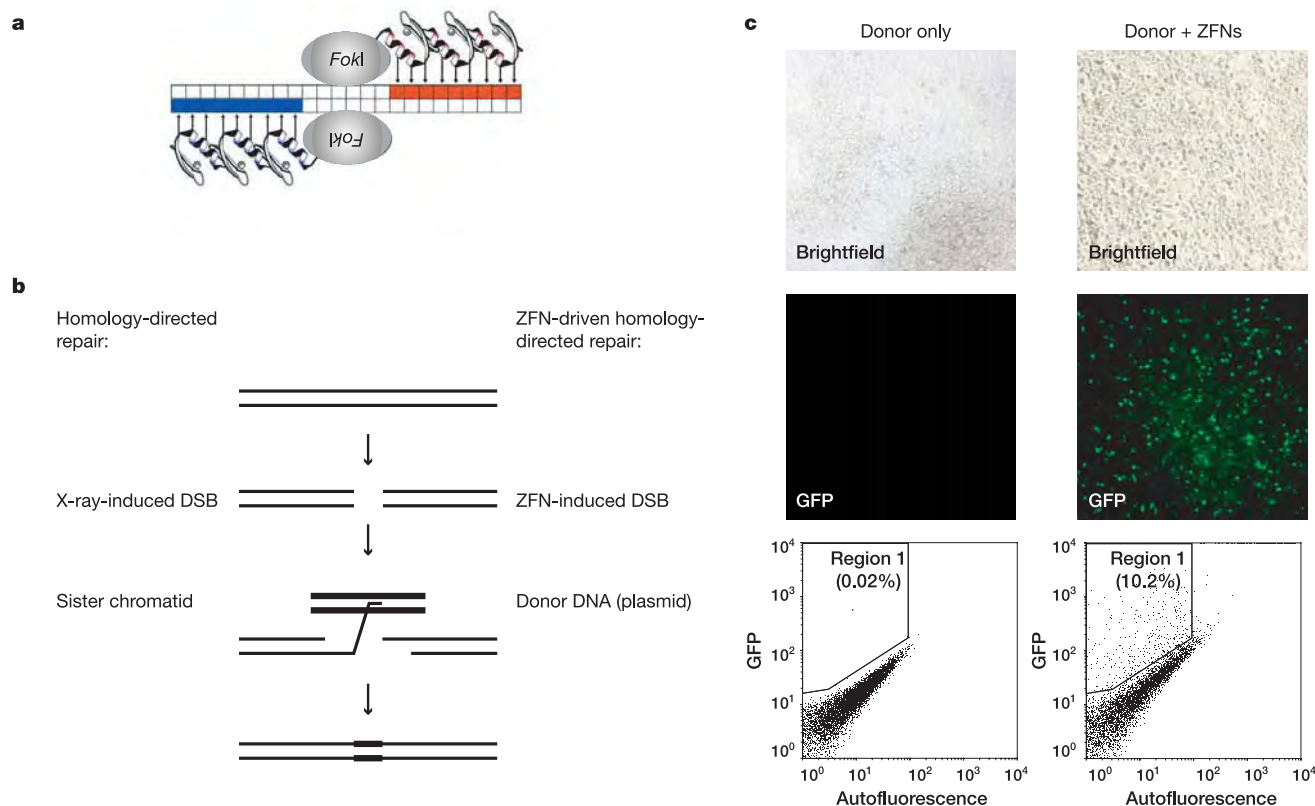


Figure 1 | Designed ZFNs enable correction of a chromosomal reporter gene in over 10% of the cells. **a**, Schematic of a DNA-bound ZFN pair. **b**, Homology-directed repair via 'short-patch gene conversion' of a DSB induced by X-rays using genetic information from a sister chromatid (left side) or of a site-specific DSB created by ZFNs, using a plasmid donor (right side). **c**, Cells carrying a mutated GFP reporter were transiently transfected with a donor plasmid carrying a fragment of wild-type GFP (left column), or the donor plasmid and the ZFNs (right column). Cells were arrested at G2/M before release and analysed by fluorescence-activated cell sorting (FACS) for GFP with GFP-positive cells bracketed in region 1.

data indicated that ZFNs could be used in a selection-free scheme to increase the rate of HR at a chromosomal reporter gene by five orders of magnitude.

Targeted alteration of DNA sequence at a human disease locus

Our ultimate goal is gene correction therapy for human disease, so we set out to determine whether ZFNs can evoke a comparable

increase in HR frequency at an endogenous gene and focused our efforts on the *IL2R γ* gene (the protein product is known as γ C), mutations in which cause X-linked SCID²⁷. Maximal HR frequency is observed when a DSB is evoked close to the mutated site²⁸, and so we used our ZFP archive²⁹ to assemble two DNA-binding domains, each containing four zinc-finger motifs and thus recognizing a total of 24 base pairs (bp) surrounding an X-linked SCID mutation hotspot

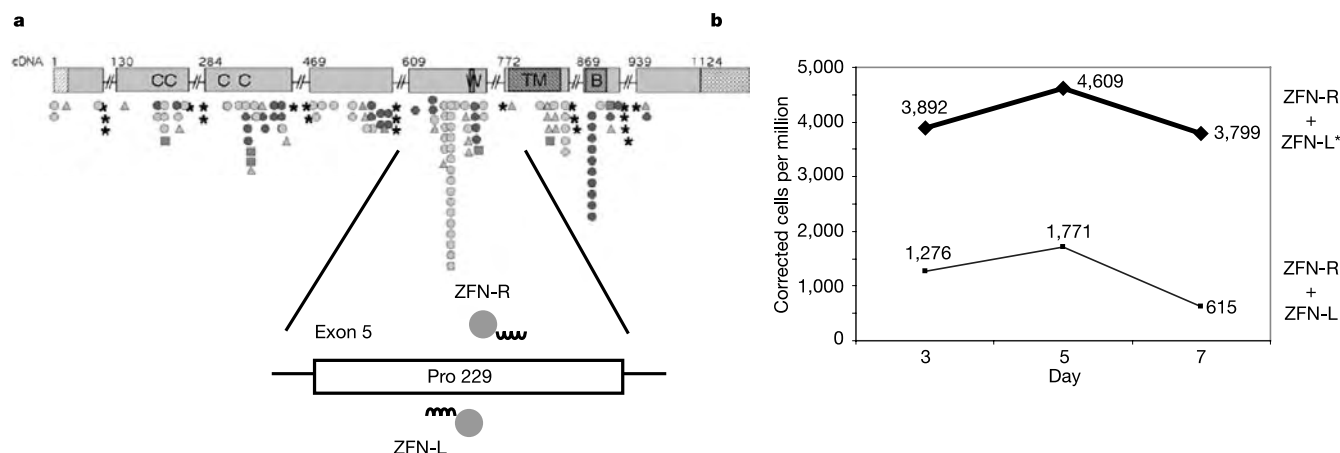
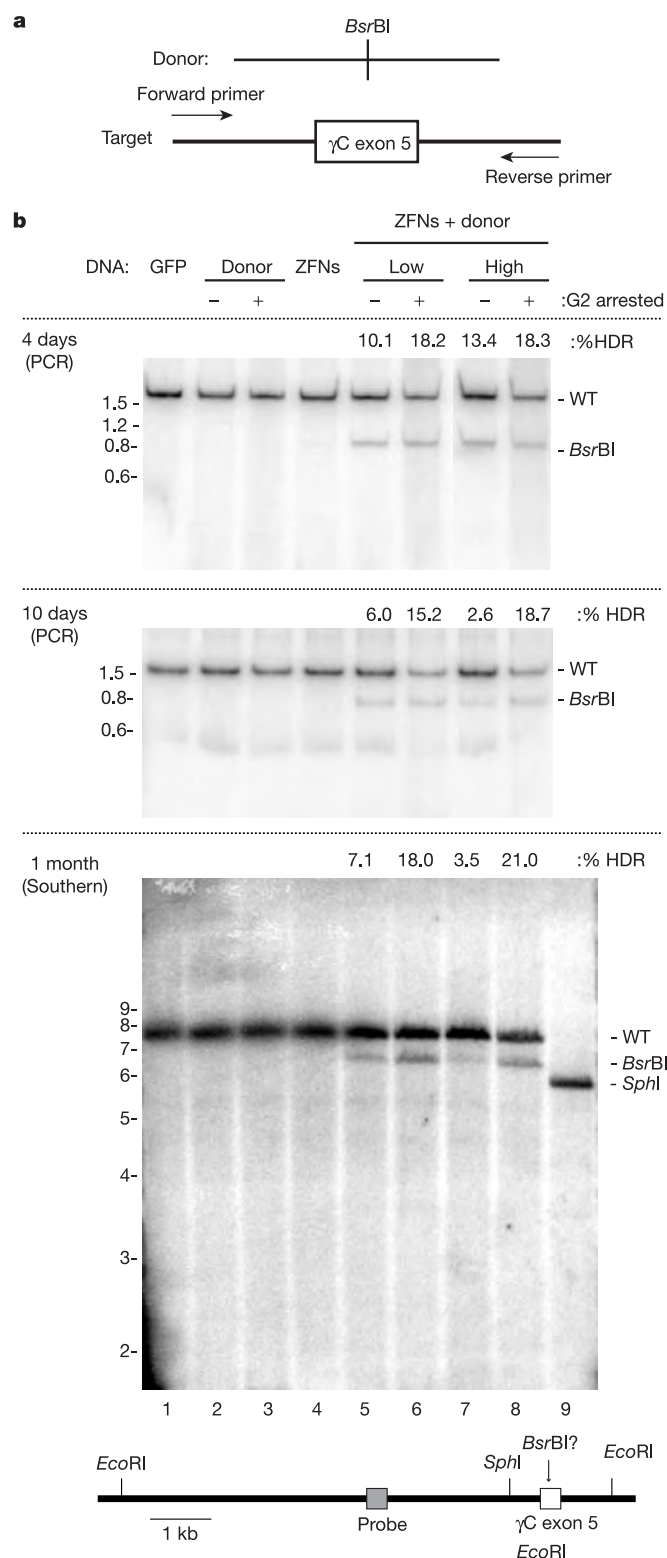


Figure 2 | Design and optimization of ZFNs directed against the X-linked SCID mutation hotspot of *IL2R γ* . **a**, Map of the human *IL2R γ* gene, with positions of disease-causing mutations (<http://genome.nih.gov/scid/IL2Rbase.shtml>)³⁰ and the ZFN binding sites annotated. **b**, ZFP optimization (upper line) markedly improves *in vivo* gene correction frequency of a chromosomal GFP reporter gene disabled by insertion of a

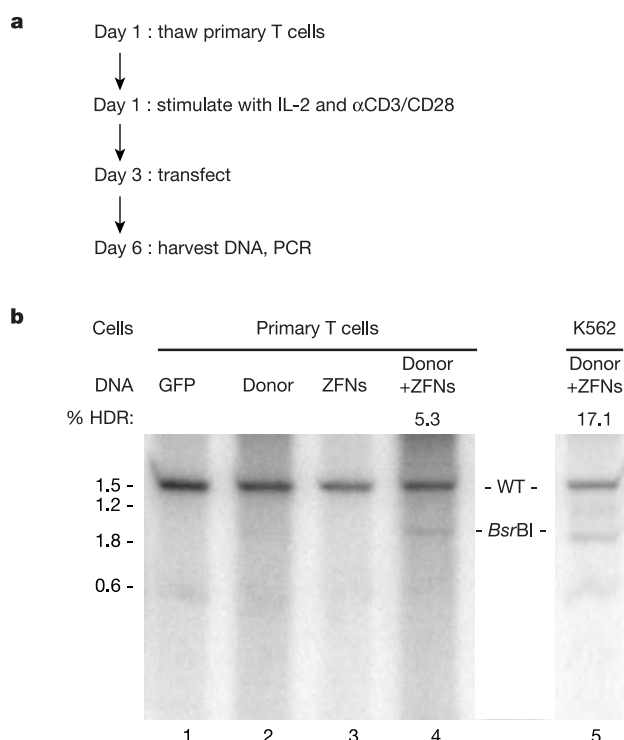
fragment of the *IL2R γ* gene carrying the ZFN recognition sites. FACS quantitation of the number of GFP-positive cells 3, 5, and 7 days post-transfection comparing archive-derived (lower line) or optimized (upper line, ZFN indicated by asterisk) ZFNs transfected along with a GFP donor plasmid.



in exon 5 (ref. 30). After initial tests for DNA binding, we optimized the ZFP–DNA interface: analysis of a large archive of designed and *in-vitro*-selected zinc-finger modules guided single amino-acid substitutions in the recognition α -helices to yield ZFPs with improved *in vitro* binding and cleavage properties (data not shown). These optimized ZFNs were ~ 5 times more potent at reporter gene correction than the original nucleases (Fig. 2b), and were well tolerated by the cells, with comparable numbers of GFP-positive cells observed at 3 and 7 days post-transfection (Fig. 2b).

We then used these optimized ZFNs to modify the sequence at the endogenous *IL2Rγ* locus. We engineered a donor plasmid containing a fragment of the *IL2Rγ* locus to carry a silent point mutation that creates a novel restriction enzyme recognition site in the exon 5 sequence (Fig. 3a). ZFN-induced HR using this donor would introduce this site into the cognate chromosomal location, and the frequency of this event can be accurately (Supplementary Fig. 1) measured by a limited-cycle polymerase chain reaction (PCR)–restriction digest assay (see Methods). Ninety-six hours after transfecting K562 cells with the donor plasmid and ZFNs we observed an HR frequency of 18% (Fig. 3b), and the conversion of the Pro 229 codon to the donor-specified form was confirmed by sequencing the *IL2Rγ* locus (data not shown). In asynchronously growing cells, HR was observed in 10% of the *IL2Rγ* alleles (Fig. 3b).

To determine whether this conversion is stable over time, we analysed genomic DNA isolated from cells at 10 days and at 1 month post-transfection with donor DNA and the ZFNs. We found that 18–21% of *IL2Rγ* loci contained the donor-specified restriction site (Fig. 3b), in both cell samples. These data eliminate the possibility of a systematic PCR-based error in the experiments, because DNA isolated at 1 month was analysed by Southern blotting. Further analysis of these DNA samples by Southern blotting failed to detect ZFN-induced donor plasmid misintegration or gross rearrangements



of the *IL2R γ* locus (Supplementary Figs 2 and 3). Identical HR frequencies at the endogenous *IL2R γ* gene were observed in cell samples harvested at 4 days and after 1 month of cell expansion (Fig. 3b, top and bottom panel), that is, no decrease in the number of corrected chromosomes was observed after a 1-month period in culture, in contrast to published work using three-finger ZFNs²⁶. To determine whether ZFN-driven targeted genome modification can be carried out in primary cells, we used the same protocol and *IL2R γ* -targeting ZFNs on human CD4⁺ T cells, and observed HR at a frequency of 5.3% (Fig. 4b)—that is, comparable to that seen earlier in transformed cells (Fig. 3b) when adjusted for the lower transfection efficiency in T cells (30%). Taken together, these data demonstrate that designed ZFNs can evoke HR to generate permanent, precise modification of the genome in 20% of the chromatids in the absence of selection.

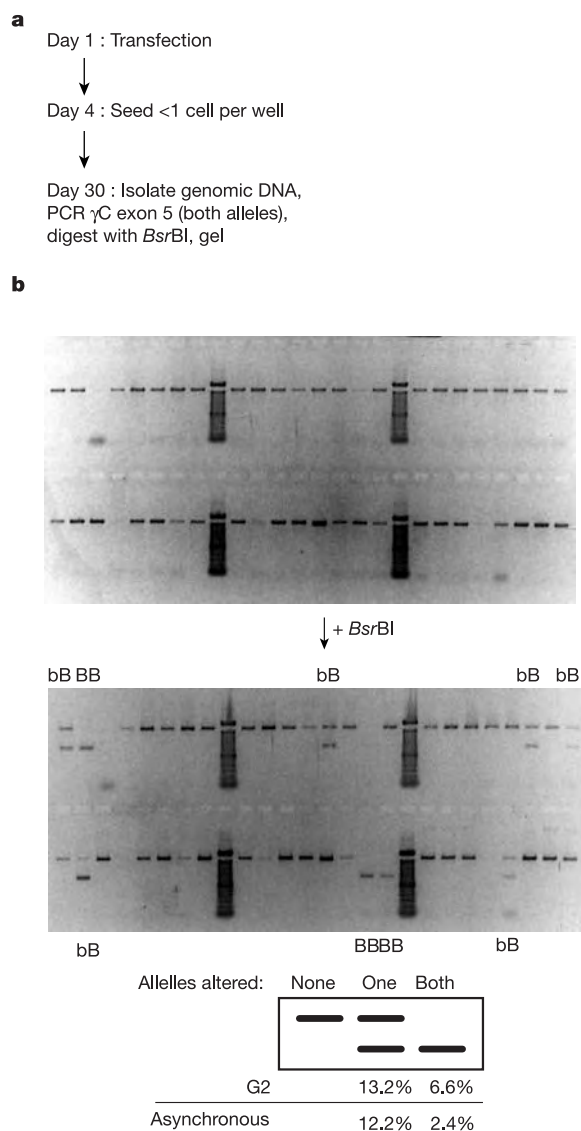


Figure 5 | High-frequency monoallelic and biallelic alteration of *IL2R γ* driven by designed ZFNs. **a**, Experimental outline. **b**, Fragments of the X chromosome carrying exon 5 of *IL2R γ* were isolated by PCR from single-cell clones after three weeks of expansion after being transfected with ZFNs and *Bsr*BI-carrying donor DNA and arrested for 30 hours in G2. Forty-eight colonies are shown here of the 96 in total that were genotyped; blank lanes correspond to samples in which DNA isolation failed for technical reasons. Clones were genotyped by digestion with *Bsr*BI (heterozygous, bB; homozygous, BB). The percentage of cells exhibiting each genotype is indicated.

Biallelic gene modification at an endogenous locus

To further characterize the ability of ZFNs to induce HR, we transfected K562 cells with the *Bsr*BI donor plasmid and ZFNs and then performed limiting dilution to isolate single clones. We genotyped exon 5 in each clone by PCR and digestion with *Bsr*BI, and found that 13.2% of the clones had converted a single allele of the *IL2R γ* gene to the donor-specified form, and that 6.6% were homozygous for that alteration (Fig. 5b). In asynchronously growing cells we found an overall HR frequency of 12.2% and a bi-allelic HR frequency of 2.4% (Supplementary Fig. 4). These data confirm the

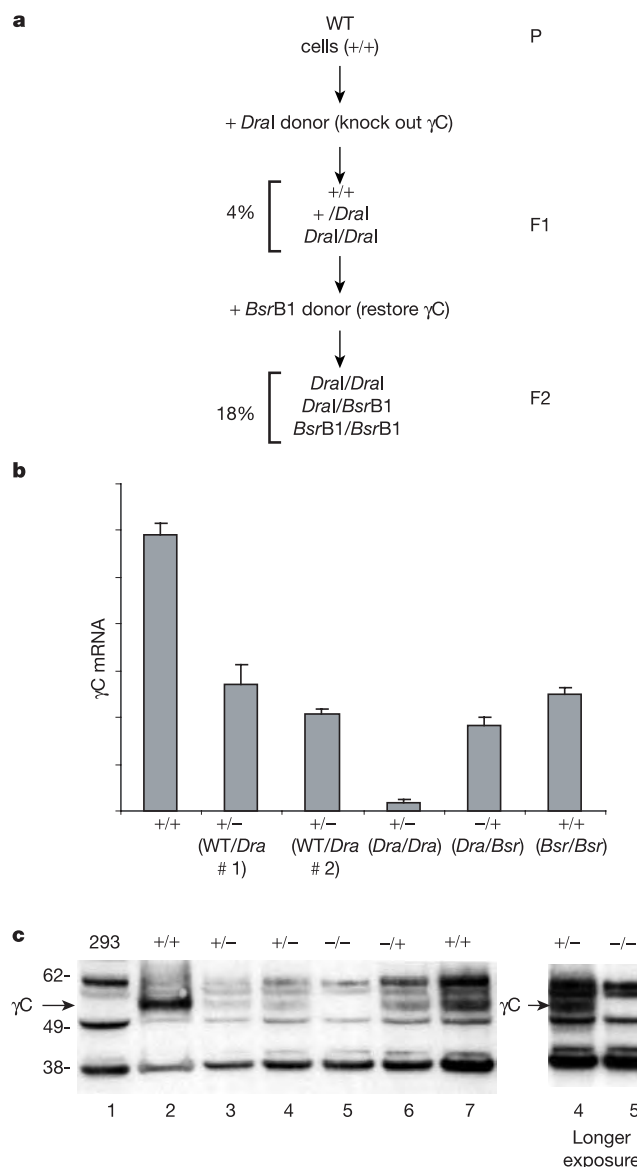


Figure 6 | Serial modification of the *IL2R γ* locus demonstrates the usefulness of ZFNs in somatic cell genetics and the potential for gene correction therapy. **a**, Experimental outline: *IL2R γ* is disabled in K562 cells using ZFNs and a donor DNA molecule carrying a frameshift and a novel restriction site (*Dra*I). Cells homozygous for the knockout allele are then reverted to a wild-type phenotype using the same ZFNs and a donor that restores the γ C open reading frame in exon 5 and creates a diagnostic restriction site (*Bsr*BI). Cell generations are shown on the right. **b**, **c**, Precise correlation between cell genotype and γ C mRNA levels as measured by quantitative real-time RT-PCR (**b**; error bars are the standard deviation in measurement of four samples) and of γ C protein detected by western blotting (**c**). Two different heterozygous clones were analysed: WT/*Dra* #1 and #2. HEK 293 cells do not express γ C. WT, wild-type.

overall frequencies found by our analysis of cell populations (Fig. 3) and further indicate that ZFNs can be used to rapidly invoke single-step, permanent biallelic modification of a specific locus in the human genome without selection.

ZFN-driven targeted alteration of mRNA and protein levels

We next demonstrated the use of ZFN-mediated HR to alter or correct the expression of an endogenous gene. We created an *IL2R γ* donor molecule that would introduce a single base pair frameshift concomitant with a *DraI* recognition site in exon 5. We transfected K562 cells with the *DraI* donor and ZFN expression plasmids, and isolated cell lines in which one or both alleles had been mutated (Fig. 6). Quantitative reverse transcription (RT)–PCR and western blot analysis showed that the resulting heterozygous cells had reduced amounts of mRNA and protein and that cells homozygous for the *DraI* site contained no detectable γ C protein or mRNA (Fig. 6b, c); the lack of mRNA is probably due to nonsense-mediated decay³¹. We next corrected this induced frameshift mutation by transfecting cells homozygous for the *DraI*-donor-derived mutation with the *BsrBI* donor and ZFNs, followed by isolation of cell clones with one or both alleles of *IL2R γ* corrected. Quantitative RT–PCR and western analysis demonstrated that these cells regained expression of both γ C mRNA and protein (Fig. 6c). RT–PCR of similarly passaged wild-type cells confirmed that the partial recovery of γ C mRNA levels in homozygous-corrected cells relative to the parental cell line (Fig. 6b) is a consequence of decreased γ C expression in the course of passaging these cells (data not shown). These experiments demonstrated that we can efficiently create cell lines carrying defined heterozygous and homozygous alterations in the genome, which lead to changes in cognate mRNA and protein levels. Furthermore, we have shown that ZFNs can efficiently correct a mutation at a locus mutated in human disease.

Discussion

The results presented here establish a general method for the rapid and permanent modification of the human genome at a specific location both in transformed and in primary cells. Building on extensive previous work^{15,32,33}, we combine high-fidelity DNA recognition by engineered C₂H₂ ZFPs and homology-directed repair of double-strand breaks to achieve HR frequencies that make feasible experimental procedures using human cells that were previously considered impractical. After only a 4-day period and in the absence of any selection for the desired event, ZFNs induce the modification of an endogenous locus in ~20% of the population, with approximately 7% of the cells becoming homozygous for the donor-specified genotype (Figs 3–5). We find these modified cells to be stable for extended periods in cell culture while transcriptionally and translationally manifesting their new genotype (Fig. 6).

The practical application of ZFN-driven gene correction in clinical settings or in basic research requires the ability to invoke a DSB at an unmodified endogenous locus with high efficiency. Work over the past 15 years has evolved the C₂H₂ class of zinc-finger DNA-binding domains into the only engineered peptide motif capable of binding to essentially any DNA sequence^{10,13,34}, and modulating the function of specific genes *in vivo*^{12,15}. A large archive of zinc-finger modules has been generated using both rational design and selection methods to provide the technical platform necessary for targeting endogenous loci^{14,15}. In agreement with the results of published work^{25,26,35} we find that both affinity and specificity are critical determinants for ZFN efficacy in driving selection-free gene correction, and of the long-term stability of the modified cell (Figs 2 and 3); for the present work, we used a large archive of both designed and selected zinc-finger modules to guide the ZFN optimization effort.

ZFNs can be used to rapidly generate cells heterozygous or homozygous for a genotype of interest. The frequency and precision of ZFN-driven HR revolutionizes the existing toolbox for mammalian somatic cell genetics and the study of gene function both in basic

research and in drug discovery efforts. Our ultimate goal is to apply the ZFN technology to the therapeutic modification of human cells in the clinic. This application will require extensive study of ZFN safety in appropriate model systems. Limitations are imposed by the need to deliver ZFN-encoding and donor DNA molecules to cells, and the potential immunogenicity of the ZFNs. These concerns are lessened in therapeutic strategies involving *ex vivo* autologous cell manipulation and monogenic diseases of haematopoiesis, including immune deficiencies and haemoglobinopathies, which lend themselves to this approach because both the appropriate target cells and the delivery methods have been extensively studied^{3,4}. Importantly, the ‘hit and run’ mechanism of ZFN action uncouples the therapeutically beneficial changes made to the genome from any need to integrate exogenous DNA, while still generating a permanently modified cell.

METHODS

ZFN and donor construct assembly. Zinc-finger proteins were designed against the coding sequence of eGFP and assembled exactly as described^{136,37}, to yield the following ZFP moieties (target gene; ZFP name; target sequence; recognition α -helices): eGFP; ZFP-L; GGGGTAGCG; RSDDLTR, QSGALAR, RSDHLR and eGFP; ZFP-R; GAAGCAGCA; QSGSLTR, QSGDLTR, QSGNLAR. Zinc-finger proteins for targeting the *IL2R γ* locus were assembled from an archive of *in vitro*-selected modules^{29,38}, assembled as described^{132,39}, and after α -helix optimization, yielded the following ZFP moieties: IL2R γ ; ZFP-R*; ACTCTGTGGAAG; RSDNLVS, RNAHRIN, RSDTLSE, ARSTRTN and IL2R γ ; ZFP-L*; AAAGCGGCTCCG; RSDTLSE, ARSTRIT, RSDLSK, QRNLKV. Assembled ZFPs were cloned in-frame as NH₂-terminal fusions to the catalytic domain of FokI^{17,25,26} into pcDNA 3.1 (Invitrogen). The donor plasmids for correcting the defective eGFP gene (see below) and for modifying the endogenous γ C locus are described in the Supplementary Information.

Gene correction in tissue culture cells. A defective eGFP reporter gene with nucleotides 229–236 relative to the start codon replaced with a stop codon and a 2-bp frameshift in the open reading frame was generated by standard PCR mutagenesis techniques, cloned into pcDNA4/TO vector (Invitrogen) and stably introduced into HEK 293 T-Rex cells (Invitrogen). A cell line with a single copy of the plasmid integrated into the genome was identified and used for all experiments. A reporter cell line carrying GFP disabled with an insertion of the ZFN-targeted stretch of *IL2R γ* was made as described²⁶; the absolute HR frequency observed when using this reporter line is lower than that seen on endogenous loci, most probably owing to the interruption of the donor–target alignment with an exogenous DNA stretch. HEK 293 T-Rex (Invitrogen) and K562 (ATCC) cells were grown according to the suppliers’ instructions and transfected with LipofectAMINE 2000 (Invitrogen) or by Nucleofection (Solution V, Program T16) (Amaxa Biosystems) according to the manufacturer’s protocol (see Supplementary Information). Gene correction of the mutated GFP reporter was measured by FACS. For analysis of gene correction at the γ C locus, genomic DNA was amplified in 20 cycles of PCR with primers that hybridize to the chromosomal *IL-2R γ* locus immediately outside of the region corresponding to the 1.5-kilobase (kb) donor sequence, and analysed by digestion with *BsrBI*, gel electrophoresis and autoradiography as described in Supplementary Information. Southern blotting (Fig. 3) was performed by probing *EcoRI*-, *BsrBI*- and *DpnI*-digested genomic DNA transferred to Nytran Plus (Schleicher and Schuell) with a 340-bp fragment of the *IL2R γ* locus labelled with α -³²P-dCTP and α -³²P-dATP (Roche) in RapidHyb buffer (Amersham Pharmacia). Quantitative RT–PCR and western blot assays for γ C were performed following standard procedures described in Supplementary Information.

Human CD4⁺ T cells (AllCells) were grown according to the supplier’s instructions. Cells were activated using 20 Units per ml IL-2 (R&D Systems) and α CD3/CD28 beads (Miltenyi Biotec, T Cell Activation/Expansion kit), and transfected with ZFN and donor constructs at an efficiency of ~30%, using an Amaxa nucleofactor according to the manufacturer’s instructions. DNA was isolated and HR at exon 5 of *IL2R γ* was measured by PCR as above, except that 25 cycles of PCR were performed.

Received 17 November 2004; accepted 18 March 2005.

Published online 3 April 2005.

1. Sedivy, J. M. & Joyner, A. L. *Gene Targeting* (Oxford Univ. Press, Oxford, 1992).
2. Yanez, R. J. & Porter, A. C. Therapeutic gene targeting. *Gene Ther.* 5, 149–159 (1998).
3. Kohn, D. B., Sadelain, M. & Glorioso, J. C. Occurrence of leukaemia following gene therapy of X-linked SCID. *Nature Rev. Cancer* 3, 477–488 (2003).

4. Persons, D. A. & Nienhuis, A. W. Gene therapy for the hemoglobin disorders. *Curr. Hematol. Rep.* **2**, 348–355 (2003).
5. Thomas, K. R., Folger, K. R. & Capecchi, M. R. High frequency targeting of genes to specific sites in the mammalian genome. *Cell* **44**, 419–428 (1986).
6. Brown, J. P., Wei, W. & Sedivy, J. M. Bypass of senescence after disruption of p21CIP1/WAF1 gene in normal diploid human fibroblasts. *Science* **277**, 831–834 (1997).
7. Bunz, F. *et al.* Targeted inactivation of p53 in human cells does not result in aneuploidy. *Cancer Res.* **62**, 1129–1133 (2002).
8. Miller, J., McLachlan, A. D. & Klug, A. Repetitive zinc-binding domains in the protein transcription factor IIIA from *Xenopus oocytes*. *EMBO J.* **4**, 1609–1614 (1985).
9. Tupler, R., Perini, G. & Green, M. R. Expressing the human genome. *Nature* **409**, 832–833 (2001).
10. Pavletich, N. P. & Pabo, C. O. Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 Å. *Science* **252**, 809–817 (1991).
11. Klug, A. Protein designs for the specific recognition of DNA. *Gene* **135**, 83–92 (1993).
12. Choo, Y., Sanchez-Garcia, I. & Klug, A. *In vivo* repression by a site-specific DNA-binding protein designed against an oncogenic sequence. *Nature* **372**, 642–645 (1994).
13. Pabo, C. O., Peisach, E. & Grant, R. A. Design and selection of novel Cys2His2 zinc finger proteins. *Annu. Rev. Biochem.* **70**, 313–340 (2001).
14. Choo, Y. & Isalan, M. Advances in zinc finger engineering. *Curr. Opin. Struct. Biol.* **10**, 411–416 (2000).
15. Jamieson, A. C., Miller, J. C. & Pabo, C. O. Drug discovery with engineered zinc-finger proteins. *Nature Rev. Drug Discov.* **2**, 361–368 (2003).
16. Kim, Y. G., Cha, J. & Chandrasegaran, S. Hybrid restriction enzymes: zinc finger fusions to Fok I cleavage domain. *Proc. Natl Acad. Sci. USA* **93**, 1156–1160 (1996).
17. Smith, J., Berg, J. M. & Chandrasegaran, S. A detailed study of the substrate specificity of a chimeric restriction enzyme. *Nucleic Acids Res.* **27**, 674–681 (1999).
18. West, S. C. Molecular views of recombination proteins and their control. *Nature Rev. Mol. Cell Biol.* **4**, 435–445 (2003).
19. Symington, L. S. Role of RAD52 epistasis group genes in homologous recombination and double-strand break repair. *Microbiol. Mol. Biol. Rev.* **66**, 630–670 (2002).
20. Paques, F. & Haber, J. E. Multiple pathways of recombination induced by double-strand breaks in *Saccharomyces cerevisiae*. *Microbiol. Mol. Biol. Rev.* **63**, 349–404 (1999).
21. Rouet, P., Smih, F. & Jasin, M. Expression of a site-specific endonuclease stimulates homologous recombination in mammalian cells. *Proc. Natl Acad. Sci. USA* **91**, 6064–6068 (1994).
22. Sargent, R. G., Brennen, M. A. & Wilson, J. H. Repair of site-specific double-strand breaks in a mammalian chromosome by homologous and illegitimate recombination. *Mol. Cell. Biol.* **17**, 267–277 (1997).
23. Chouluka, A., Perrin, A., Dujon, B. & Nicolas, J. F. Induction of homologous recombination in mammalian chromosomes by using the I-SceI system of *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* **15**, 1968–1973 (1995).
24. Bibikova, M. *et al.* Stimulation of homologous recombination through targeted cleavage by chimeric nucleases. *Mol. Cell. Biol.* **21**, 289–297 (2001).
25. Bibikova, M., Beumer, K., Trautman, J. K. & Carroll, D. Enhancing gene targeting with designed zinc finger nucleases. *Science* **300**, 764 (2003).
26. Porteus, M. H. & Baltimore, D. Chimeric nucleases stimulate gene targeting in human cells. *Science* **300**, 763 (2003).
27. Cavazzana-Calvo, M. *et al.* Gene therapy of human severe combined immunodeficiency (SCID)-X1 disease. *Science* **288**, 669–672 (2000).
28. Elliott, B., Richardson, C., Winderbaum, J., Nickoloff, J. A. & Jasin, M. Gene conversion tracts from double-strand break repair in mammalian cells. *Mol. Cell. Biol.* **18**, 93–101 (1998).
29. Moore, M., Klug, A. & Choo, Y. Improved DNA binding specificity from polyzinc finger peptides by using strings of two-finger units. *Proc. Natl Acad. Sci. USA* **98**, 1437–1441 (2001).
30. Buckley, R. H. Primary immunodeficiency diseases due to defects in lymphocytes. *N. Engl. J. Med.* **343**, 1313–1324 (2000).
31. Schell, T., Kulozik, A. E. & Hentze, M. W. Integration of splicing, transport and translation to achieve mRNA quality control by the nonsense-mediated decay pathway. *Genome Biol.* **3**, 1–6 (2002).
32. Isalan, M. & Choo, Y. Rapid, high-throughput engineering of sequence-specific zinc finger DNA-binding proteins. *Methods Enzymol.* **340**, 593–609 (2001).
33. Wilson, J. H. Pointing fingers at the limiting step in gene targeting. *Nature Biotechnol.* **21**, 759–760 (2003).
34. Choo, Y. & Klug, A. Physical basis of a protein-DNA recognition code. *Curr. Opin. Struct. Biol.* **7**, 117–125 (1997).
35. Bibikova, M., Golic, M., Golic, K. G. & Carroll, D. Targeted chromosomal cleavage and mutagenesis in *Drosophila* using zinc-finger nucleases. *Genetics* **161**, 1169–1175 (2002).
36. Zhang, L. *et al.* Synthetic zinc finger transcription factor action at an endogenous chromosomal site. Activation of the human erythropoietin gene. *J. Biol. Chem.* **275**, 33850–33860 (2000).
37. Liu, P. Q. *et al.* Regulation of an endogenous locus using a panel of designed zinc finger proteins targeted to accessible chromatin regions. Activation of vascular endothelial growth factor A. *J. Biol. Chem.* **276**, 11323–11334 (2001).
38. Isalan, M., Klug, A. & Choo, Y. A rapid, generally applicable method to engineer zinc fingers illustrated by targeting the HIV-1 promoter. *Nature Biotechnol.* **19**, 656–660 (2001).
39. Tan, S. *et al.* Zinc-finger protein-targeted gene regulation: genomewide single-gene specificity. *Proc. Natl Acad. Sci. USA* **100**, 11997–12002 (2003).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We are grateful to C. Case for his initial discussions with M.H.P. regarding this project. We thank Sangamo's production group for technical support; S. Brennan, C. Dent, D. Kohn, Y. Marahrens, T. Martin, C. Pabo and P. Sung for suggestions and discussions; and A. Klug for comments on the manuscript. We also thank E. Lanphier for encouragement and support.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare competing financial interests: details accompany the paper on www.nature.com/nature. Correspondence should be addressed to M.C.H. (mholmes@sangamo.com) or M.H.P. (matthew.porteus@UTSouthwestern.edu); requests for materials should be addressed to M.C.H.

LETTERS

A resolved outflow of matter from a brown dwarf

Emma T. Whelan¹, Thomas P. Ray¹, Francesca Bacciotti², Antonella Natta², Leonardo Testi²
& Sofia Randich²

The birth of stars involves not only accretion but also, counter-intuitively, the expulsion of matter in the form of highly supersonic outflows^{1,2}. Although this phenomenon has been seen in young stars, a fundamental question is whether it also occurs among newborn brown dwarfs: these are the so-called 'failed stars', with masses between stars and planets, that never manage to reach temperatures high enough for normal hydrogen fusion to occur³. Recently, evidence for accretion in young brown dwarfs has mounted^{4–6}, and their spectra show lines that are suggestive of outflows^{7–9}. Here we report spectro-astrometric data that spatially resolve an outflow from a brown dwarf. The outflow's characteristics appear similar to, but on a smaller scale than, outflows from normal young stars. This result suggests that the outflow mechanism is universal, and perhaps relevant even to the formation of planets.

The nearby ρ Ophiuchi cloud (at a distance of 125 pc; ref. 10) is an excellent example of a stellar nursery. It was as part of a near-infrared survey of this region¹¹ that ρ Oph 102 was first detected. Its mass is given as $60M_J$ (ref. 7), where $1M_J$ equals the mass of Jupiter or 0.001 solar masses ($0.001M_\odot$). This places it firmly within the brown dwarf mass range ($0.013M_\odot \leq M_{BD} \leq 0.075M_\odot$)³, and indeed this object has been spectroscopically confirmed to be a brown dwarf⁷. There is also strong evidence for the presence of an accretion disk⁴ and, in fact, the accretion rate (derived from its H α profile)⁷ is estimated to be approximately $10^{-9}M_\odot\text{yr}^{-1}$. Finally, it has been noted that the spectrum of ρ Oph 102 contains a number of forbidden emission lines, suggesting an outflow⁷.

Tracing an outflow from a young star is done using a variety of techniques depending upon wavelength. When the star itself is sufficiently evolved to be optically visible—that is, much of the surrounding natal gas and dust has been driven away—the outflow can be followed almost right back to the star through its permitted and forbidden line emission¹. Nearest the source, the light from the collimated outflow is most intense, giving rise to the so-called 'micro-jets' observed, for example, from T Tauri stars, the precursors to stars like our Sun¹². In extreme cases, the outflow is observed only very close ($\leq 1''$) to the source¹³: that is, within the typical 'seeing' disk for ground-based telescopes. It is under such circumstances that the technique of spectro-astrometry comes to the fore (see Methods and Supplementary Information).

Our high resolution echelle observations (see Methods and Fig. 1) of ρ Oph 102 show its forbidden emission lines; for example, [O I] at wavelengths of 6,300 and 6,363 Å and [S II] at 6,731 Å, are moderately blueshifted (radial velocity $V_r \approx -45\text{ km s}^{-1}$). Such radial velocities are very similar to those seen in outflows from classical T Tauri stars, and might be expected of an outflow from a young brown dwarf, as its escape velocity is similar¹⁴. We note also the absence of a redshifted outflow component. This is again typical of T Tauri stars, and is interpreted in terms of an obscuring disk that hides the redshifted component from our view¹³. We can thus immediately infer the

presence of a disk around this brown dwarf (see below for a limit on its size). Such a conclusion is also in line with the mid-infrared excess seen from this object⁴. (Note that all velocities are with respect to the systemic velocity of the brown dwarf. The latter was derived from the Li photospheric absorption line at 6,708 Å, and equals $7 \pm 8\text{ km s}^{-1}$ in the Local Standard of Rest frame.)

Another indication that an outflow is present comes from a cursory examination of the H α line (Fig. 2). Its profile is clearly asymmetrical, that is, the blueshifted wing of the line appears to be absorbed in a P Cygni-like fashion. We note however that we do not observe a classical P Cygni profile, that is, one that dips below the continuum. Such a profile is in any event a rare occurrence even amongst T Tauri stars¹⁵.

If we are dealing with a scaled-down version of the outflow phenomenon seen in T Tauri stars, then we expect the centre of emission in forbidden lines to be spatially offset from the continuum, that is, the brown dwarf. This offset is due to the fact that, for a collimated outflow, such lines are quenched close to the star once the electron density becomes high enough¹³. In the case of outflows from T Tauri stars, typical offsets of 30–75 AU ($0.2''$ – $0.5''$ at 150 pc) are seen, for example, in the [O I] doublet at 6,300 and 6,363 Å and the [S II] doublet at 6,717 and 6,731 Å (ref. 13). If we assume that brown dwarf outflows have similar opening angles and velocities to those from T Tauri stars, then the point at which the critical density is reached may naively be taken to scale with $\dot{M}_{\text{jet}}^{1/2}$, where $\dot{M}_{\text{jet}}$ is the jet's mass loss rate (see Supplementary Information). Assuming that the latter depends linearly on the accretion rate, we would expect typical spatial offsets to be 3–10 times smaller in brown dwarf outflows in comparison to those from T Tauri stars.

Spectro-astrometric (emission centroid offset versus velocity) plots are shown in Fig. 1 for the [O I] doublet, H α and the [S II] line at 6,731 Å. Because of high electron densities close to the brown dwarf, the [S II] line at 6,717 Å was too weak to provide a usable spectro-astrometric signal. Here the spatial offsets were measured after continuum subtraction (see Methods and Supplementary Information for details).

We now consider the main results from the various offset versus velocity plots. First, the centroids of all the measurable forbidden lines are displaced to the south, that is, they have negative offsets with respect to the continuum. These offsets reach a maximum of $0.08''$ – $0.1''$ at a blueshifted velocity of about -40 km s^{-1} . We have already noted the absence of any corresponding redshifted emission and that this suggests the presence of an obscuring disk. The scale of the blueshifted offset would imply a minimum (projected) disk radius of $0.1''$ ($\geq 15\text{ AU}$ at the distance of the ρ Ophiuchi cloud) in order to hide any redshifted component.

Second, there is no clear spatial offset in H α , even though its higher signal to noise ratio potentially allows us to measure even smaller offsets than observed in the forbidden lines. This is in agreement with the idea that most of the H α emission arises from accretion⁷ on much smaller scales than are being probed here. It is

¹Dublin Institute for Advanced Studies, 5 Merrion Square, Dublin 2, Ireland. ²Osservatorio Astrofisico di Arcetri, INAF, Largo E. Fermi 5, I-50125 Firenze, Italy.

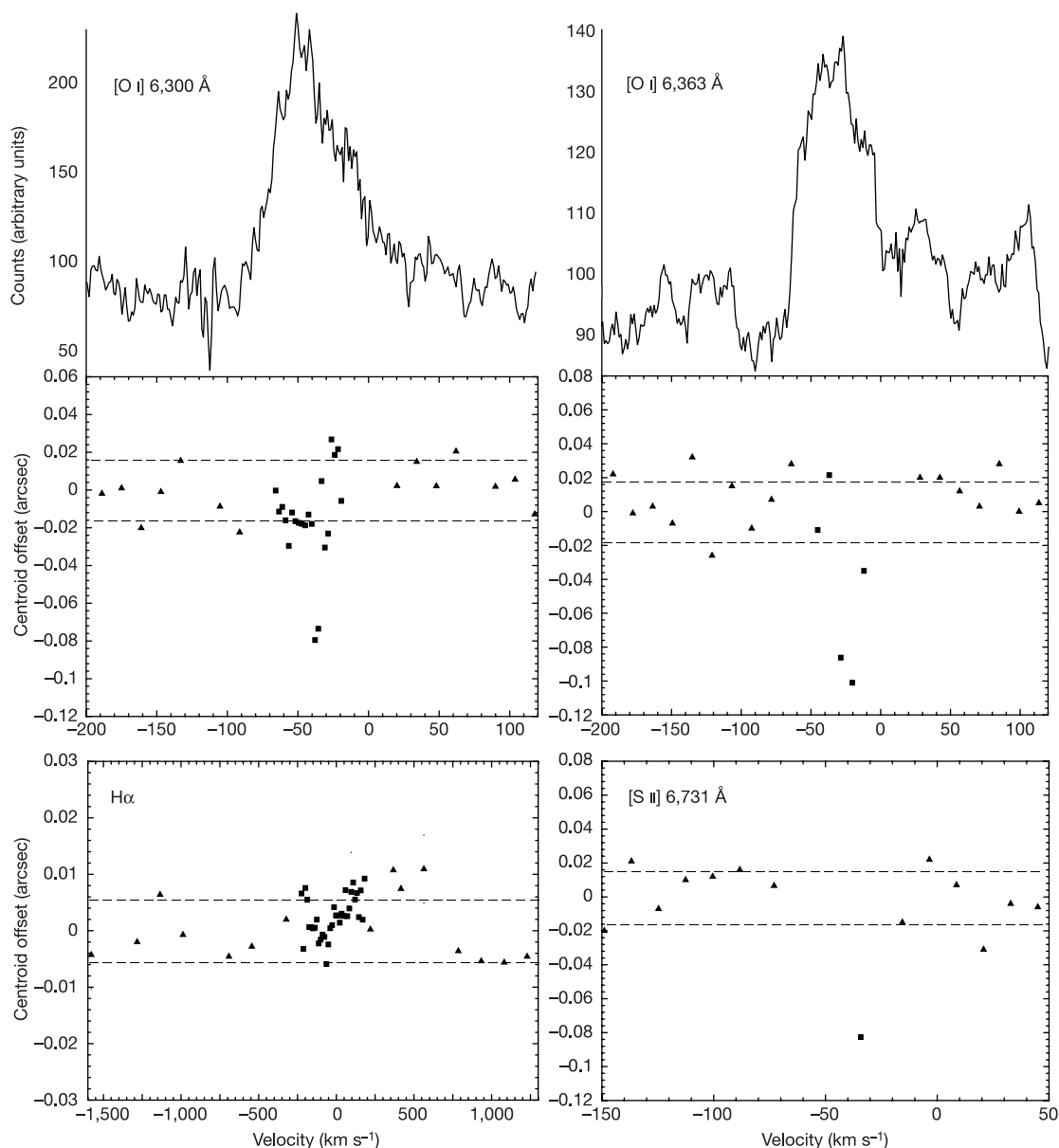


Figure 1 | Resolving the outflow from ρ Oph 102. Line profiles (top row) and spectro-astrometric plots (middle row) for the [O I] doublet at 6,300 and 6,363 Å, and spectro-astrometric plots (bottom row) for the H α and the 6,731 Å line of [S II]. The [O I] night sky lines at -9 km s^{-1} are subtracted. No offset measurements were made in their vicinity (thus accounting for the data point gaps). Continuum and line offset points are

represented by triangles and squares, respectively. Velocities are systemic, and offsets are in the north–south direction with negative offsets to the south. Dashed lines delineate the $\pm 1\sigma$ error envelope. For H α , note the much smaller offset scale. The [S II] line is blueshifted to around -40 km s^{-1} .

also worth pointing out that if, as is almost certainly the case, the blue-ward dip in the H α profile is caused by P Cygni-like absorption from an outflowing wind in front of the star, no offset should be expected.

Third, both the line profiles and the spectro-astrometric signatures are very similar (albeit on somewhat smaller scales) to what is seen in T Tauri stars with ‘micro-jets’. In particular, the observed velocities and offsets in the various forbidden lines are within the range we would expect for a collimated outflow from a brown dwarf (see also Supplementary Information). We suggest that direct imaging (using, for example, Fabry-Perot systems) of this and other candidate brown dwarf outflows should now be attempted. Such observations will, however, be very challenging, even with large telescopes, because of the expected faintness of the outflow.

METHODS

Spectro-astrometry. Conceptually the principles of spectro-astrometry are easy to understand. The profile of a star is smeared by atmospheric turbulence to appear gaussian (at least to a first approximation) rather than point-like. Whereas the width of the profile is determined by the so-called seeing, how accurately we can determine the centroid of emission is, in theory for fixed seeing, limited only by the strength of the observed signal to noise ratio. Increasing the total number of detected photons increases the positional, or astrometric, accuracy, so that, in principle, milliarcsecond precision is possible with very large ground based telescopes^{16–18}.

Consider now a long-slit spectrum of a close binary system consisting of two virtually identical stars. We will assume that the slit is orientated along the same position angle as the binary. (We note that strictly this is not necessary: it is only necessary that the slit is not orthogonal.) If the separation of the binary is considerably less than the seeing, the profile of the system in the spatial direction

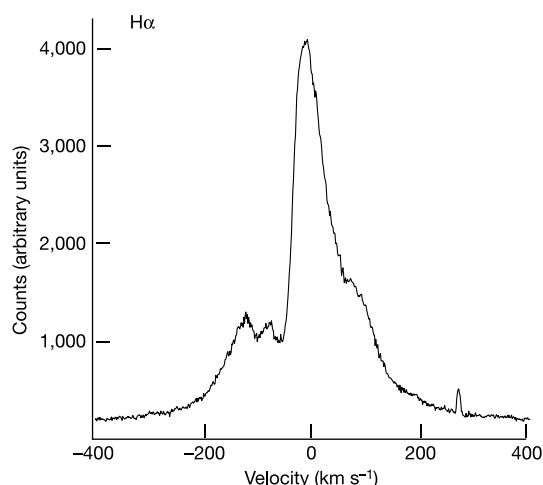


Figure 2 | The outflow signature in the H α profile of ρ Oph 102. The P Cygni-like dip in the line profile is a strong signature of outflow activity. H α emission from the brown dwarf is absorbed as it passes through material moving outwards along our line of sight. Because this material is moving towards us, the dip is on the blueward side of the line. Classical T Tauri stars are strong H α emitters, and P Cygni H α profiles originally confirmed that such protostars drive outflows. The dip in the H α emission of ρ Oph 102 is at approximately the outflow radial velocity determined from the forbidden lines.

will consist of a single gaussian—that is, the system is unresolved and the centroid of emission will lie exactly between the two components. Now suppose that one of the two stars differs slightly from the other in being a strong H α emitter; in such a case, the emission centroid will shift towards that star in the spectrum at the position of the H α line. In this way it is possible to resolve certain types of binaries with separations well within the seeing limit¹⁹. In the case of a jet (pure emission line region) plus star (continuum source), one can go further and interpolate the continuum across a line, thereby allowing its contribution to be removed. It is then possible to measure separately the spatial centroid of the pure emission line region and determine its offset with respect to the continuum, that is, the parent star. Moreover, as the line can be emitted over a range of wavelengths, owing to the Doppler effect, it may also be feasible to recover spatio-kinematic information. For example, if the jet is bipolar, that is, it has oppositely directed blue- and redshifted flows from the source; the emission centroid of the red and blue wings of the line will be displaced to opposite sides of the continuum centre.

The detailed method by which we measure offsets can briefly be described as follows. First, the centroid of the continuum emission in the spatial direction is determined using a one-dimensional gaussian fit. The line of such centroids, in the dispersion direction but excluding any region where emission lines are present, is then fitted with a second-order polynomial, over a range of typically 200–300 Å. In this way, instrumental curvature and tilting, with a characteristic frequency many times larger than the width of any line, is determined. The fit, to the centre of the continuum, is then subtracted from the actual measured centroids, leaving residuals that are evenly scattered about the abscissa (that is, the fit defines the zero offset line). The continuum data points shown in Fig. 1 are thus the residuals. Finally, the two-dimensional fit to the continuum, broadened to take account of the point spread function, is subtracted from the emission lines. Any emission line offsets are then measured.

The accuracy (in arcseconds) of the method is set by the error in the centroid of the gaussian fit, which depends on the seeing and the number of detected photons, N . Formally, the error is given by $\text{Seeing}/[2(2 \ln 2)^{1/2}N^{1/2}]$, assuming that photon noise is the only source of noise. N , of course, is a function of the binning and the spatial sampling (pixel width). This explains why, for example, we can achieve a higher spectro-astrometric accuracy with a bright line, such as H α , than a weak one, for example, the [S II] line at 6,731 Å. In some cases, it is necessary to bin up a weak line in the dispersion direction, as we have done to varying degrees for the [O I] doublet and the [S II] line at 6,731 Å, to achieve sufficient signal to noise ratio. Note that we sometimes use different binning factors for the continuum, in comparison with the line, so as to achieve a similar

signal to noise ratio in both. This allows us to have comparable offset errors in both components, and to define the common 1σ error lines shown in Fig. 1. As can be seen from the plots, the typical limiting offset that we can measure in the spatial direction (3σ) is around 30 mas. This corresponds to 4.5 AU at the distance to the ρ Ophiuchi cloud.

Echelle spectroscopy. The high resolution spectra of ρ Oph 102 were taken with the UV-visual Echelle Spectrograph (UVES) on the European Southern Observatory's 8 m Kueyen Telescope, one of the telescopes in the Very Large Telescope (VLT) suite, in May 2003. A total of three 45 min exposures of the target were made, together with a series of flats and biases as well as an observation of an arc lamp for wavelength calibration. The slit was orientated north-south and had a width of 1" while the seeing was 0.65". The central wavelength was set at 580 nm, giving a spectral range of 450–680 nm. Only the red part of the spectrum from 580 to 680 nm, however, was analysed. The pixel scale was 0.182" and the spectral resolution $R = 40,000$. The data were reduced using standard Image Reduction and Analysis Facility (IRAF) routines.

Received 23 September 2004; accepted 23 March 2005.

1. Eislöffel, J., Mundt, R., Ray, T. P. & Rodríguez, L. F. in *Protostars and Planets IV* (eds Mannings, V., Boss, A. P. & Russell, S. S.) 815–840 (University of Arizona Press, Tucson, 2000).
2. Königl, A. & Pudritz, R. E. in *Protostars and Planets IV* (eds Mannings, V., Boss, A. P. & Russell, S. S.) 759–787 (University of Arizona Press, Tucson, 2000).
3. Basri, G. Observations of brown dwarfs. *Annu. Rev. Astron. Astrophys.* **38**, 485–519 (2000).
4. Natta, A. *et al.* Exploring brown dwarf disks in ρ Ophiuchi. *Astron. Astrophys.* **393**, 597–609 (2002).
5. Pascucci, I., Apai, D., Henning, T. & Dullemond, C. P. The first detailed look at a brown dwarf disk. *Astrophys. J.* **590**, L111–L114 (2003).
6. Jayawardhana, R., Mohanty, S. & Basri, G. Evidence for a T Tauri phase in young brown dwarfs. *Astrophys. J.* **592**, 282–287 (2003).
7. Natta, A. *et al.* Accretion in brown dwarfs: An infrared view. *Astron. Astrophys.* **424**, 603–612 (2004).
8. Comerón, F., Fernández, M., Baraffe, I., Neuhauser, R. & Kaas, A. A. New low-mass members of the Lupus 3 Dark Cloud: Further indications of pre-main sequence evolution strongly affected by accretion. *Astron. Astrophys.* **406**, 1001–1017 (2003).
9. Fernández, M. & Comerón, F. Intense accretion and mass loss of a very low mass young stellar object. *Astron. Astrophys.* **380**, 264–276 (2001).
10. de Geus, E. J., de Zeeuw, P. T. & Lub, J. Physical parameters of stars in the Scorpio-Centaurus OB Association. *Astron. Astrophys.* **216**, 44–61 (1989).
11. Greene, T. P. & Young, E. T. Near-infrared observations of young stellar objects in the ρ Ophiuchi Dark Cloud. *Astrophys. J.* **395**, 516–528 (1992).
12. Dougados, C., Cabrit, S., Lavalley, C. & Ménard, F. T Tauri star microjets resolved by adaptive optics. *Astron. Astrophys.* **357**, L61–L64 (2000).
13. Hirth, G. A., Mundt, R. & Solf, J. Spatial and kinematic properties of the forbidden emission line region of T Tauri stars. *Astron. Astrophys. Suppl.* **126**, 437–469 (1997).
14. Masciadri, E. & Raga, A. C. Looking for outflows from brown dwarfs. *Astrophys. J.* **615**, 850–854 (2004).
15. Muzerolle, J., Calvet, N. & Hartmann, L. Emission-line diagnostics of T Tauri magnetospheric accretion. II. Improved model tests and insights into accretion physics. *Astrophys. J.* **550**, 944–961 (2001).
16. Takami, M., Bailey, J., Gledhill, T. M., Chrysostomou, A. & Hough, J. H. Circumstellar structure of RU Lupi down to AU scales. *Mon. Not. R. Astron. Soc.* **323**, 177–187 (2001).
17. Whelan, E. T., Ray, T. P. & Davis, C. J. Paschen beta emission as a tracer of outflow activity from T-Tauri stars, as compared to optical forbidden emission. *Astron. Astrophys.* **417**, 247–261 (2004).
18. Takami, M., Bailey, J. & Chrysostomou, A. A spectro-astrometric study of southern pre-main sequence stars. Binaries, outflows, and disc structure down to AU scales. *Astron. Astrophys.* **397**, 675–691 (2003).
19. Bailey, J. Detection of pre-main-sequence binaries using spectro-astrometry. *Mon. Not. R. Astron. Soc.* **301**, 161–167 (1998).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This work was supported in part by Science Foundation Ireland and the JETSET Marie Curie research training network.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to E.T.W. (ewhelan@cp.dias.ie).

Ultrafast non-thermal control of magnetization by instantaneous photomagnetic pulses

A. V. Kimel¹, A. Kirilyuk¹, P. A. Usachev², R. V. Pisarev², A. M. Balbashov³ & Th. Rasing¹

The demand for ever-increasing density of information storage and speed of manipulation has triggered an intense search for ways to control the magnetization of a medium by means other than magnetic fields^{1–5}. Recent experiments on laser-induced demagnetization^{6–8} and spin reorientation⁹ use ultrafast lasers as a means to manipulate magnetization, accessing timescales of a picosecond or less. However, in all these cases the observed magnetic excitation is the result of optical absorption followed by a rapid temperature increase. This thermal origin of spin excitation considerably limits potential applications because the repetition frequency is limited by the cooling time¹⁰. Here we demonstrate that circularly polarized femtosecond laser pulses can be used to non-thermally excite and coherently control the spin dynamics in magnets by way of the inverse Faraday effect. Such a photomagnetic interaction is instantaneous and is limited in time by the pulse width (~ 200 fs in our experiment). Our finding thus reveals an alternative mechanism of ultrafast coherent spin control, and offers prospects for applications of ultrafast lasers in magnetic devices.

The interaction of light with magnetized media is manifested in various magneto-optical phenomena. A good example is the Faraday effect, observed as a rotation of the polarization plane of light transmitted through a magnetic medium¹¹:

$$\alpha_F = \frac{\chi}{n} \mathbf{M} \cdot \mathbf{k} \quad (1)$$

where α_F is the specific Faraday rotation, $\mathbf{M}$ is the magnetization, n is the refractive index, $\mathbf{k}$ is the wave vector of light, and χ is the magneto-optical susceptibility, which is a scalar value in isotropic media^{12,13}. Various devices, such as magneto-optical isolators and modulators, make use of large values of Faraday rotation in transparent magnetic compounds.

Much less known is the inverse Faraday effect, where high-intensity laser radiation induces a static magnetization $\mathbf{M}(0)$:

$$\mathbf{M}(0) = \frac{\chi}{16\pi} [\mathbf{E}(\omega) \times \mathbf{E}^*(\omega)] \quad (2)$$

where $\mathbf{E}(\omega)$ and $\mathbf{E}^*(\omega)$ are the electric field of the light wave and its complex conjugate, respectively^{13–16}. It follows from equation (2) that circularly polarized light at frequency ω should induce a magnetization along the wave vector $\mathbf{k}$. Note that symmetry considerations of equation (2) indicate equivalence between photoexcitation by circularly polarized light and action of an external magnetic field. Moreover, right- and left-handed circularly polarized waves should induce magnetizations of opposite sign.

Equations (1) and (2) show that both these phenomena are determined by the same magneto-optical susceptibility χ (refs 14, 15). In particular, in the case of the inverse Faraday effect, χ is the ratio between the induced magnetization and the laser intensity. Therefore, optical control of magnetization is expected to

be most efficient in materials with high values of the Faraday rotation per unit magnetization. Another important property of the susceptibility χ is that it has no symmetry restrictions and is thus allowed in all media, regardless of their crystallographic and magnetic structures. Moreover, the inverse Faraday effect does not require absorption, and is based on a Raman-like coherent optical scattering process. This has the important consequence that the effect of light on the magnetization is non-thermal and can be considered as instantaneous because it takes place on a femtosecond timescale. Indeed, if one stimulates an optical transition into a virtual state with a strong spin-orbit interaction, the following relaxation into the ground state may be accompanied by spin switching and re-emission of a photon with a fixed phase shift and lower energy with respect to that of the incident photon. In magnetically ordered materials, this process is known as excitation of magnons by light¹⁷. Recent theoretical work has indicated the possibility of laser-induced spin reversal on a femtosecond timescale¹⁸. However, the experimental demonstration of such non-thermal ultrafast optical control of magnetization has remained an intriguing challenge until now.

The material of choice for our study was dysprosium orthoferrite

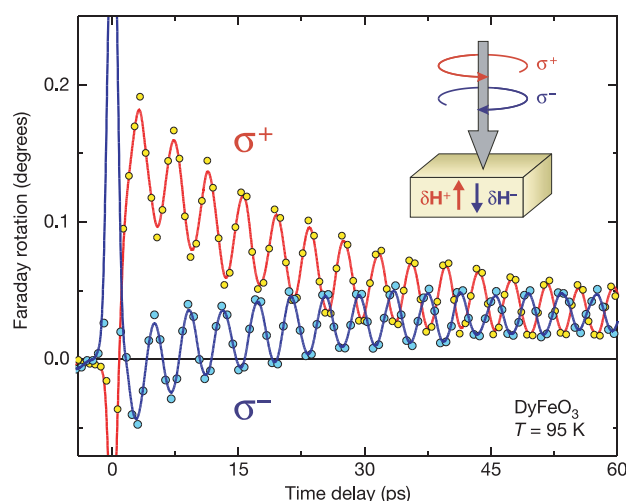


Figure 1 | Magnetic excitations in DyFeO₃ probed by the magneto-optical Faraday effect. Two processes can be distinguished: (1) instantaneous changes of the Faraday effect due to the photoexcitation of Fe ions and relaxation back to the high spin ground state $S = 5/2$; (2) oscillations of the Fe spins around their equilibrium direction with an approximately 5 ps period. The circularly polarized pumps of opposite helicities excite oscillations of opposite phase. Inset shows the geometry of the experiment. Vectors δH^+ and δH^- represent the effective magnetic fields induced by right-handed σ^+ and left-handed σ^- circularly polarized pumps, respectively.

¹IMM, Radboud University Nijmegen, 6525 ED Nijmegen, The Netherlands. ²IoFFE Physico-Technical Institute, 194021 St Petersburg, Russia. ³Moscow Power Engineering Institute, 111250 Moscow, Russia.

DyFeO₃, which belongs to the group of rare-earth orthoferrites¹⁹. This material crystallizes in an orthorhombic perovskite-type structure with four molecular units per unit cell, with a space-group symmetry D_{2h}^{16} ($Pbnm$). The spins of the dysprosium ions are not ordered above 4 K, being in a paramagnetic state. The spins of the Fe³⁺ ions ($3d^5$, ground state $^6A_{1g}$, $S = 5/2$) are coupled antiferromagnetically by isotropic exchange. The Dzyaloshinskii–Moriya interaction^{20,21} leads to a slight canting of opposite spins with an angle of about 0.5°, giving rise to a spontaneous magnetization $M_s = 8$ G. Despite the small magnetization, this material exhibits a giant Faraday rotation of about $3,000^\circ \text{cm}^{-1}$ owing to its strong spin–orbit interaction²². Thus non-thermal effects of light on the spontaneous magnetization are expected to be large in this material.

The studied DyFeO₃ samples were prepared from X-ray oriented boules grown by a floating zone method. They were cut perpendicular to the z and x crystal axes and were 60 μm thick. For the detection of the optically induced magnetization, we used the direct magneto-optical Faraday effect. Figure 1 shows the temporal evolution of the Faraday rotation in a z -cut sample for two circularly polarized pump pulses of opposite helicities. On the scale of 60 ps one can clearly distinguish two different processes that start after excitation with a pump pulse. At zero time delay, instantaneous changes of the Faraday rotation are observed that result from the excitation of virtual and real transitions in the Fe³⁺ ions from the high spin ground state $S = 5/2$. The instantaneous changes of the Faraday rotation are followed by oscillations with a frequency of about 200 GHz, which can clearly be assigned to oscillations of the magnetization. It is seen from Fig. 1 that the helicity of the pump light controls the sign of the photo-induced magnetization. This observation unambiguously indicates that the coupling between spins and photons in DyFeO₃ is direct, because the phase of the spin oscillations is given by the sign of the angular momentum of the exciting photon.

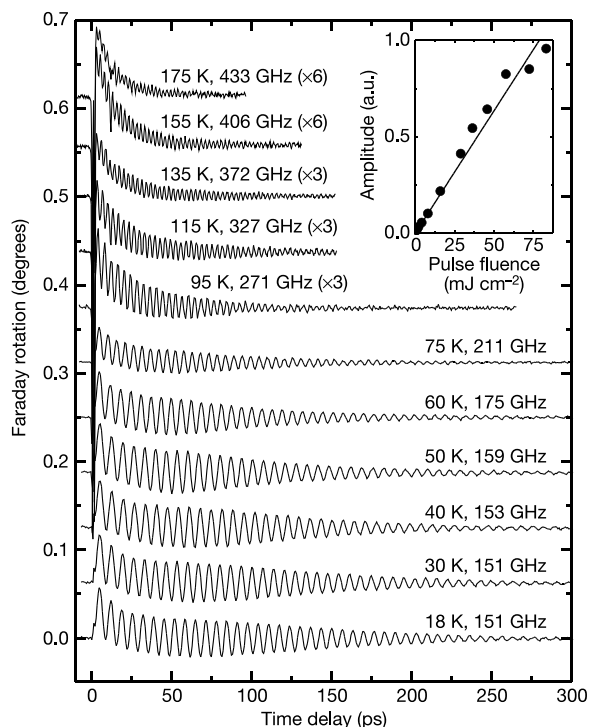


Figure 2 | Excitation of the spin oscillations in DyFeO₃ measured at different temperatures in the range between 20 K and 170 K. In order to exclude effects not relevant to magnetic excitations, the difference between the signals for right- and left-handed circularly polarized pump pulses is plotted. Every new curve is shifted from the previous one along the vertical axis over 0.06° . Inset shows the amplitude of the spin oscillations as a function of pump fluence.

Figure 2 shows the difference between the Faraday rotations induced by right- and left-handed circularly polarized pump light in the z -cut sample for the temperature range between 20 K and 175 K. It is seen that an increase of the temperature results in an increase of the frequency of the oscillations up to 450 GHz at 175 K, while the amplitude of the oscillation decreases. This behaviour is in excellent agreement with previous Raman experiments in DyFeO₃ (refs 23–25). The damping of the oscillations in the range of 200 ps is due to magnon scattering on phonons and spins of dysprosium ions. The highest value of the amplitude of the photo-induced oscillations is observed between 20 K and 50 K. The amplitude of the oscillations corresponds to a photo-induced magnetization $M = M_s/16$, where M_s is the saturation magnetization. This ratio is obtained from hysteresis measurements in a static magnetic field that show that the saturated Faraday rotation in a single domain z -cut sample is equal to 1° .

From Figs 1 and 2 one can distinguish not only oscillations but also an exponential decay of the equilibrium point on a timescale of about 100 ps. This can be explained by the photo-induced change of the equilibrium orientation of the magnetization and subsequent decay of the equilibrium orientation to the initial state.

Although in principle the effect of optically induced magnetization does not require the absorption of photons, laser control of the spontaneous magnetization and the excitation of coherent spin oscillations is equivalent to photoexcitation of magnons and thus requires some energy. Such photoexcitation of magnons can occur via a process similar to Raman scattering. The inset in Fig. 2 shows the amplitude of the photoexcited spin oscillations as a function of the pump intensity. The linearity of this dependence indicates that the photoexcitation of magnons is a one-photon process. Note that

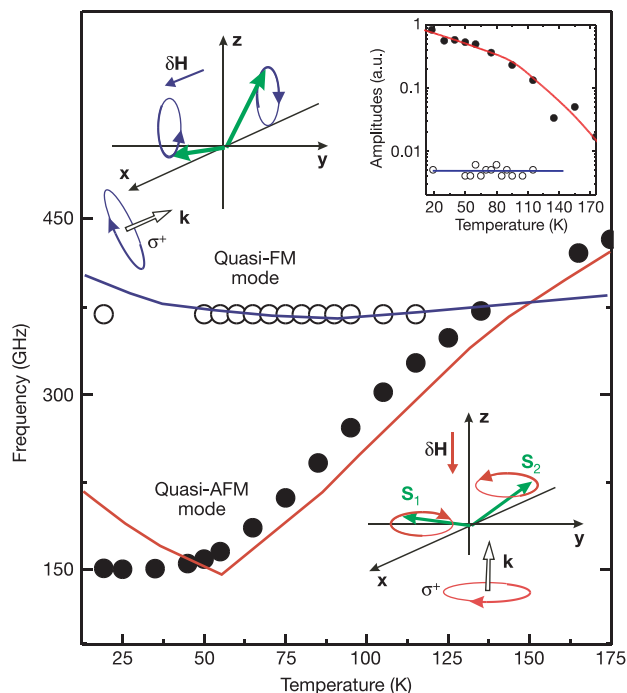


Figure 3 | Temperature dependence of the frequencies of the observed spin oscillations. Filled and open circles show the frequencies of the excited oscillations for laser pulses propagating along the z axis and x axis, respectively. Red and blue lines show the frequency of the quasi-antiferromagnetic (quasi-AFM) and the quasi-ferromagnetic (quasi-FM) resonance modes from refs 23–25. Top right inset shows the temperature dependence of the oscillation amplitudes. Top left and bottom right insets are respectively schematic representations of the quasi-FM and quasi-AFM modes of the spin resonance. Vectors $\delta\mathbf{H}$ show the directions of the instantaneous magnetic field that is equivalent to the photoexcitation.

extrapolation of the intensity dependence shows that the photo-induced effect on the magnetization would reach the saturation value of M_S at a pump fluence of about 500 mJ cm^{-2} . The effect of such a 200 fs laser pulse on the magnetic system is equivalent to the application of a magnetic field pulse of about 5 T. According to our measurements, the absorption in DyFeO_3 in the near-infrared spectral range is of the order of $100\text{--}200 \text{ cm}^{-1}$. Given this low value of the absorption, a photoexcitation of 500 mJ cm^{-2} is still below the damage threshold of DyFeO_3 and is thus quite feasible, given a sample of high optical quality.

Owing to the strong anisotropy of the magnetic susceptibility in DyFeO_3 , magnetic fields in different directions should trigger different types of spin oscillations (Fig. 3). A magnetic field pulse directed along the z axis excites oscillations that correspond to the quasi-antiferromagnetic resonance mode, whereas a field pulse along the x axis will excite the quasi-ferromagnetic resonance mode²³. These predictions are in excellent agreement with the experimentally observed temperature dependence of the frequency of the oscillations for z -cut and x -cut samples, which closely resemble the temperature dependence for the quasi-antiferromagnetic and the upper quasi-ferromagnetic resonance mode in DyFeO_3 , respectively (see Fig. 3). All these observations unambiguously show that an ultrashort laser pulse acts on the ensemble of strongly correlated spins as a magnetic field pulse directed along the wave vector of the photons. Simple estimates show that such optical pulses are equivalent to magnetic field pulses with an amplitude of 0.3 T and a full-width at half-maximum of about 200 fs.

Note that the application of a static external magnetic field up to 0.05 T in a direction parallel to the wave vector of light only resulted in a slight change of the frequency (about 1%), again confirming that the effective photo-induced field is dominating the dynamics.

We have demonstrated that with circularly polarized femtosecond laser pulses one can purely optically and thus non-thermally excite and coherently control spin oscillations in the weak ferromagnet DyFeO_3 . Such optical pulses are shown to be equivalent to 200 fs magnetic field pulses up to 5 T. In view of the great variety of magnetic materials, the direct effect of light on spontaneous magnetization in other materials and at higher temperatures is foreseen. Our findings open new insights into the understanding of ultrafast magnetic excitation and, regarding recent progress in the development of compact ultrafast lasers²⁶, may provide new prospects for applications of ultrafast photomagnetic phenomena.

METHODS

The measurements were performed in a pump and probe configuration at a photon energy of 1.55 eV using amplified 200 fs pulses from a Ti:sapphire laser at a repetition rate of 1 kHz. The pump beam was circularly polarized, while the probe beam had linear polarization. The intensity ratio between the pump and probe pulses was about 100. Both beams were focused on the sample to a spot diameter of about $200 \mu\text{m}$ for the pump and somewhat smaller for the probe beam. The pump fluence on the sample was around 30 mJ cm^{-2} . The measurements were done in a cold finger cryostat where the temperature could be stabilized in the range 15–300 K with a precision better than 0.5 K.

Received 25 November 2004; accepted 15 March 2005.

Published online 25 May 2005.

- Asamitsu, A., Tomioka, Y., Kuwahara, H. & Tokura, Y. Current switching of resistive states in magnetoresistive manganites. *Nature* **388**, 50–52 (1997).

- Ohno, H. *et al.* Electric-field control of ferromagnetism. *Nature* **408**, 944–946 (2000).
- Kato, Y., Myers, R. C., Gossard, A. C. & Awschalom, D. D. Coherent spin manipulation without magnetic fields in strained semiconductors. *Nature* **427**, 50–53 (2004).
- Yamanouchi, M., Chiba, D., Matsukura, F. & Ohno, H. Current-induced domain-wall switching in a ferromagnetic semiconductor structure. *Nature* **428**, 539–542 (2004).
- Lottermoser, Th. *et al.* Magnetic phase control by an electric field. *Nature* **430**, 541–544 (2004).
- Beaurepaire, E., Merle, J.-C., Daunois, A. & Bigot, J.-Y. Ultrafast spin dynamics in ferromagnetic nickel. *Phys. Rev. Lett.* **76**, 4250–4253 (1996).
- Hohlfeld, J., Matthias, E., Knorren, R. & Bennemann, K. H. Nonequilibrium magnetization dynamics of nickel. *Phys. Rev. Lett.* **78**, 4861–4864 (1997).
- Koopmans, B., van Kampen, M., Kohlhepp, J. T. & de Jonge, W. J. M. Ultrafast magneto-optics in nickel: magnetism or optics? *Phys. Rev. Lett.* **85**, 844–847 (2000).
- Kimel, A. V., Kirilyuk, A., Tsvetkov, A., Pisarev, R. V. & Rasing, Th. Laser-induced ultrafast spin reorientation in the antiferromagnet TmFeO_3 . *Nature* **429**, 850–853 (2004).
- Hohlfeld, J. *et al.* Fast magnetization reversal of GdFeCo induced by femtosecond laser pulses. *Phys. Rev. B* **65**, 012413 (2002).
- Faraday, M. On the magnetization of light and the illumination of magnetic lines of force. *Phil. Trans. R. Soc. Lond.* **136**, 104–123 (1846).
- Landau, L. D. & Lifshitz, E. M. *Electrodynamics of Continuous Media* (Pergamon, Oxford, 1984).
- Pitaevskii, L. P. Electric forces in a transparent dispersive medium. *Sov. Phys. JETP* **12**, 1008–1013 (1961).
- van der Ziel, J. P., Pershan, P. S. & Malmstrom, L. D. Optically-induced magnetization resulting from the inverse Faraday effect. *Phys. Rev. Lett.* **15**, 190–193 (1965).
- Pershan, P. S., van der Ziel, J. P. & Malmstrom, L. D. Theoretical discussion of inverse Faraday effect, Raman scattering and related phenomena. *Phys. Rev.* **143**, 574–583 (1966).
- Awschalom, D. D., Warnock, J. & von Molnar, S. Low-temperature magnetic spectroscopy of a dilute magnetic semiconductor. *Phys. Rev. Lett.* **58**, 812–815 (1987).
- Shen, Y. R. & Bloembergen, N. Interaction between light waves and spin waves. *Phys. Rev.* **143**, 372–384 (1960).
- Gomez-Abal, R., Ney, O., Satitkovitchai, K. & Hübner, W. All-optical subpicosecond magnetic switching in $\text{NiO}(001)$. *Phys. Rev. Lett.* **92**, 227402 (2004).
- Wijn, H. P. J. (ed.) Numerical Data and Functional Relationships. In *Landolt-Börnstein New Series Group III*, Vol. 27, f3, 125–134 (Springer, Berlin, 1981).
- Dzyaloshinskii, I. E. Thermodynamic theory of weak ferromagnetism in antiferromagnetic substances. *Sov. Phys. JETP* **5**, 1259–1272 (1957).
- Moriya, T. Anisotropic superexchange interaction and weak ferromagnetism. *Phys. Rev.* **120**, 91–98 (1960).
- Zvezdin, A. K. & Kotov, V. A. *Modern Magneto-optics and Magneto-optical Materials* (IOP, Bristol, 1997).
- White, R. M., Nemanich, R. J. & Herring, C. Light scattering from magnetic excitations in orthoferrites. *Phys. Rev. B* **25**, 1822–1836 (1982).
- Balbashov, A. M., Volkov, A. A., Lebedev, S. P., Mukhin, A. A. & Prokhorov, A. S. High-frequency magnetic properties of dysprosium orthoferrite. *Sov. Phys. JETP* **61**, 573–586 (1985).
- Koshizuka, N. & Hayashi, K. Raman scattering from magnon excitations in RFeO_3 . *J. Phys. Soc. Jpn* **57**, 4418–4428 (1988).
- Keller, U. Recent developments in compact ultrafast lasers. *Nature* **424**, 831–838 (2003).

Acknowledgements We thank V. N. Gridnev for discussions, and T. Toonen and A. van Etteger for help with measurements. This work was partially supported by the European IST network SPINOSA, the RTN network DYNAMICS, the Russian Foundation for Basic Research (RFBR), de Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO), and de Stichting voor Fundamenteel Onderzoek der Materie (FOM).

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to T.R. (th.rasing@science.ru.nl).

LETTERS

Field regulation of single-molecule conductivity by a charged surface atom

Paul G. Piva^{1,2}, Gino A. DiLabio², Jason L. Pitters², Janik Zikovsky¹, Moh'd Rezek^{1,2}, Stanislav Dogel¹, Werner A. Hofer³ & Robert A. Wolkow^{1,2}

Electrical transport through molecules has been much studied since it was proposed¹ that individual molecules might behave like basic electronic devices, and intriguing single-molecule electronic effects have been demonstrated^{2,3}. But because transport properties are sensitive to structural variations on the atomic scale^{4–7}, further progress calls for detailed knowledge of how the functional properties of molecules depend on structural features. The characterization of two-terminal structures has become increasingly robust and reproducible^{8–12}, and for some systems detailed structural characterization of molecules on electrodes or insulators is available^{13–17}. Here we present scanning tunnelling microscopy observations and classical electrostatic and quantum mechanical modelling results that show that the electrostatic field emanating from a fixed point charge regulates the conductivity of nearby substrate-bound molecules. We find that the onset of molecular conduction is shifted by changing the charge state of a silicon surface atom, or by varying the spatial relationship between the molecule and that charged centre. Because the shifting results in conductivity changes of substantial magnitude, these effects are easily observed at room temperature.

Figure 1a shows a scanning tunnelling microscope (STM) image of the H-terminated Si(100) surface of a highly n-type doped (As, $7 \times 10^{19} \text{ cm}^{-3}$) crystal obtained at 20 °C (ref. 18); the bright bar feature is a line of styrene-derived molecules. Such molecular lines grow on silicon surfaces according to a 'self-directed' process^{18,19} that automatically juxtaposes molecules in an ordered contiguous fashion and places at the end of the line a dangling bond, seen in the image at the left end of the line. The bright circular feature just below the line is a second dangling bond.

The sequence of images and cross-sections presented in Fig. 1 indicate that the 'slope effect'—the decreasing apparent height of molecules with increasing distance from the dangling bond—is somehow related to the dangling bond. At lower imaging bias (Fig. 1b, c) the molecules nearest to the dangling bond appear prematurely heightened, as if experiencing a built-in offset voltage. Molecules most distant from the dangling bond show a voltage–height response that is largely unperturbed. At larger imaging voltages (as in Fig. 1a) those distant molecules appear as high as the molecules nearest the dangling bond. As we establish below, the essence of the field-regulated molecular conduction effect described here is a shifting of molecular energy levels under the influence of the electrostatic potential emanating from a charged dangling bond. The distinct onset behaviour displayed by π -bond containing molecules causes relatively small shifts in imaging voltage to lead to pronounced changes in molecule-mediated conduction.

To understand these effects, we consider the charge state of dangling bonds and the electrostatic interactions of dangling bonds and other electronic energy levels under the influence of the field of

the STM tip. At the dangling-bond density we studied, on a highly n-doped crystal, the dangling bonds will be negatively charged²⁰. The essential features of dangling bonds have been known for approximately 50 years (ref. 21): they have the capacity to be variably charged while maintaining an energetic position within the bandgap, allowing dangling bonds to act as 'pinning' centres. On an n-type crystal, and increasingly for higher dopant concentrations, dangling bonds take on electrons; they thus become negatively charged and cause bands to bend upwards. For highly n-doped material, the Fermi level is near the conduction-band edge, substantial charge moves to the surface and band bending is of the order of 0.5 eV. For intrinsic (undoped) or low-doped material, the Fermi level is near the mid-gap point. In such cases a relatively small amount of charge is placed at the surface and band bending is also small.

The charge state of a dangling bond can be determined only if dopant concentration as well as the effect of an externally imposed electric field are known, as recently demonstrated²². We extended the approach to treat (using the FEMLAB 3 application, COMSOL, Inc., Burlington, Massachusetts) our particular materials, dopant densities, surface states (dangling bonds), and applied fields, finding that the conditions relevant to Fig. 1 cause the dangling bonds to be negatively charged. This occurs because on a very highly (degenerately) n-doped crystal, dangling bonds are effectively connected to the bulk of the crystal. In contrast, a mid-gap state on a low-doped crystal is virtually disconnected from its surroundings and cannot substantially source or sink current²² and, as a result, the dangling bond does not become negative during imaging. Figure 2 presents an image of a relatively low-doped (10^{16} cm^{-3}) n-type silicon sample. The slope effect is absent. In all instances, the local Fermi level contributes to determining the occupancy of the dangling bond and therefore biasing the tip-sample system relative to a reference potential could modulate the dangling-bond charge state.

Further evidence that the slope effect does not exist when the charge at the dangling bond is eliminated is presented in Fig. 3. Figure 3a shows two molecular lines, each terminated by a charged dangling bond and exhibiting a pronounced slope effect. When a 'TEMPO' radical, which bonds to Si dangling bonds²³, is attached to each of the dangling bonds, the charged dangling bond is eliminated and the slope effect lost (Fig. 3b). Figure 3c shows that the dangling bonds are regenerated when the TEMPO molecules are removed via a tip-induced desorption process²³, with regeneration of the negative dangling bonds restoring the slope effect. The loss and regeneration of slope is also evident in the height profiles in Fig. 3d. These observations clearly demonstrate chemical modulation (on/off) of molecular conductivity—a chemical gating action. This effect might lead to electrically monitored single-molecule-detection schemes. A variation of the procedure could be used to annihilate a dangling bond and create a new dangling bond in another position, allowing

¹Department of Physics, 534 Avadh Bhatia Physics Lab, University of Alberta, Edmonton, Alberta T6G 2J1, Canada. ²National Institute for Nanotechnology, National Research Council of Canada, W6-010 ECERF, 9107-116th Street, Edmonton, Alberta T6G 2V4, Canada. ³Surface Science Research Centre, University of Liverpool, Liverpool L69 3BX, UK.

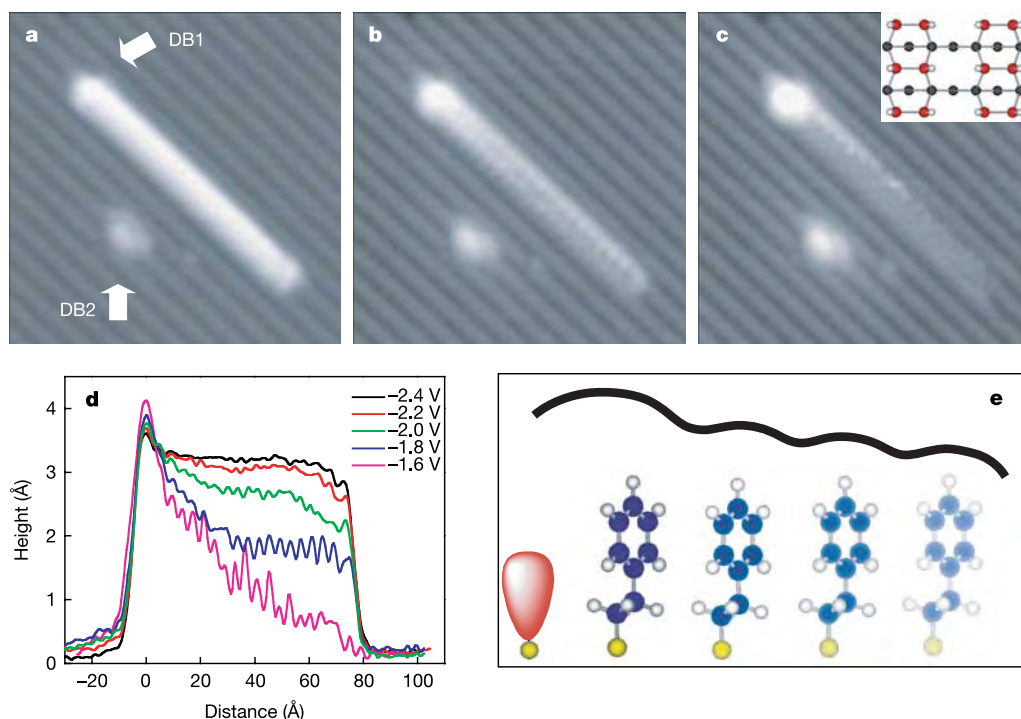


Figure 1 | Visualization of the electrostatic potential emanating from a point source. **a**, STM image of highly n-type doped H-Si(100), with negatively charged dangling bonds labelled DB1 and DB2. The prominent white bar is a line of surface-bound molecules. At increased sample bias of -2.4 V, molecular π -states are 'turned on', causing molecules to appear bright (topographically elevated) and of nearly constant height across the line. **b**, At an intermediate bias of -1.8 V, molecules appear darker, increasingly so at greater distances from DB1. **c**, In the absence of a negative dangling bond all molecules would appear dark at -1.6 V, but the molecules nearest to the dangling bond remain prominent. Molecules near the dangling bond experience a greater effective tip-sample bias due to the negatively charged dangling bond's electrostatic potential. Inset, a schematic of the silicon surface studied. **d**, Cross-sectional occupied-state height

profiles taken along the molecular line for given sample bias voltages. The effect of DB2 is particularly evident as a broad maximum in the -2.0 -V cross-section. **e**, Schematic showing the spatial relationship between the dangling bond (red paraboloid), styrene-derived molecules, and silicon substrate. Yellow, blue, and white spheres represent silicon, carbon and hydrogen atoms, respectively. The fading represents the diminishing appearance of the molecules due to reduced conduction, as seen in the experimental images. The black line is a representation of the height profile of the image. The dark line shows the sloping topographic envelope observed under low-bias conditions resulting from the negatively charged dangling bond. Images and linescan data were acquired at a constant tunnel current of 40 pA. Image areas are $10\text{ nm} \times 10\text{ nm}$.

us to explore different geometric relationships between the charge centre and the molecules.

Quantum-mechanical calculations were used to model charging and level-shifting effects. A 2×1 H-Si cluster consisting of 250 silicon atoms was constructed and the silicon back-bonds were terminated by hydrogen atoms. All atom positions were energy-optimized (surface atoms were frozen in lattice positions) using the AM1 method²⁴. A radical cluster was generated by the removal of a hydrogen atom from a surface site. An anionic cluster was generated by adding an electron to the radical cluster and shifting the silicon atom with the dangling bond to a position about $0.4\text{ \AA}$ above the other surface silicon atoms, in accordance with the results of full geometry optimizations on smaller anionic clusters. Structures of styrene lines were optimized using the HCTH407 (ref. 25)/CEP-31G (ref. 26) level of theory. All of the molecules were constrained during the optimizations to have identical structures and be perpendicular to the surface. Single-point energies were obtained using the PBE (ref. 27)/CEP-31G as implemented in the 'Gaussian' package²⁸.

The left side of Fig. 4 shows that the highest-energy molecular π -type state occurs at about 0.7 V below the valence band edge and is localized near the negative dangling bond. The summed charge densities²⁹ (right side of Fig. 4) illustrate that this localization results in the slope effect. At larger voltages, the molecular states tend to be localized farther from the dangling bond. The additional charge density centred on more distant molecules reduces the slope. Ultimately, the sum of the molecular charge densities in an energy window of 1.5 V shows that the slope effect is essentially eliminated.

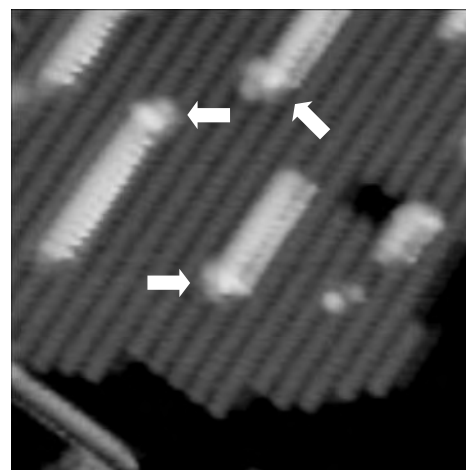


Figure 2 | Absence of charge field effects on low-doped silicon. The image shows that styrene lines on low-doped n-type H-Si(100) have no slopes. The dangling bonds (indicated by arrows) are neutral under the imaging conditions used (sample bias -2.0 V and 80 pA), and no significant height perturbation is observed along the molecular lines. Image area is $15\text{ nm} \times 15\text{ nm}$.

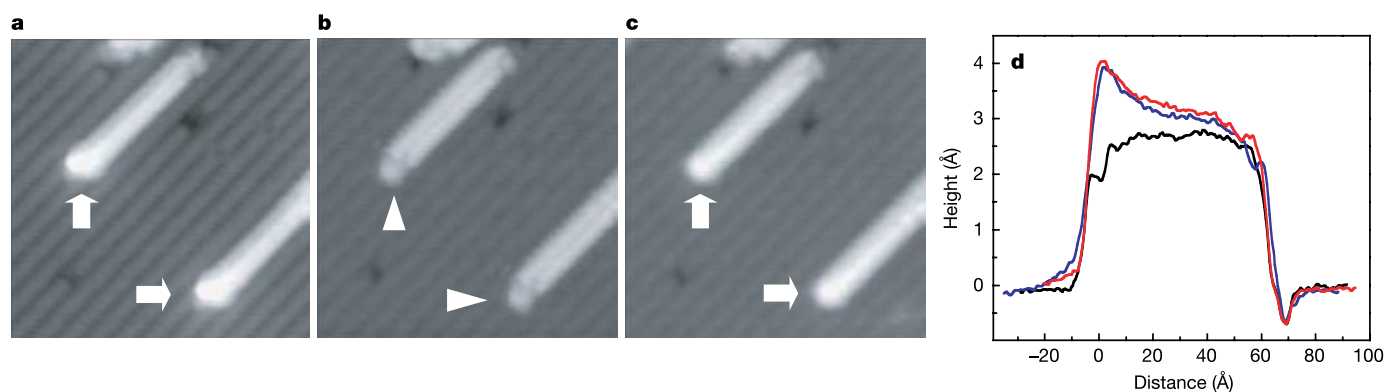


Figure 3 | Reversible modification of dangling bonds. The ‘capping’ and ‘uncapping’ of dangling bonds correlates with the disappearance and reappearance of the slope effect, as illustrated in the $10\text{ nm} \times 10\text{ nm}$ STM images acquired at -1.9 V and 50 pA . **a**, Sloping styrene lines with a dangling bond at the end of each line, indicated by arrows. **b**, One TEMPO (2,2,6,6-tetramethyl-1-piperidinyloxy) molecule reacted at each dangling bond as

indicated with wedges. Charge, and therefore slope, are absent. **c**, TEMPO molecules are removed by scanning at -3 V . The charged dangling bonds, indicated by arrows, are regenerated and the slope reappears. **d**, Profiles of styrene lines from the upper left corners of panels **a**, **b** and **c** are shown in blue, black and red, respectively.

This agrees with the STM observations of molecules nearest the dangling bond appearing to ‘turn on’ at lower-magnitude imaging voltages.

The essence of the field-regulation effect is made clear by the current–voltage spectra shown in Fig. 5a: the closer a molecule is to the charged dangling bond, the lower the sample bias required to observe the current onset. Figure 5b shows a plot of calculated π -state eigenvalues as a function of inverse distance from a localized charge centre. The linear relationship obtained is consistent with the

electrostatic origin of the effect—that is, a Stark shift³⁰.

Obvious extensions of this study include pre-programming the conduction threshold energy through molecular design and tuning the strength of the electrostatic coupling. Moreover, different electrical test configurations and variations in dangling-bond–molecule configurations can be explored. In fact, close examination of Fig. 1 illustrates an example of the latter effect: the influence of the second dangling bond is evident as a broad maximum in the -2-V cross-sectional curve.

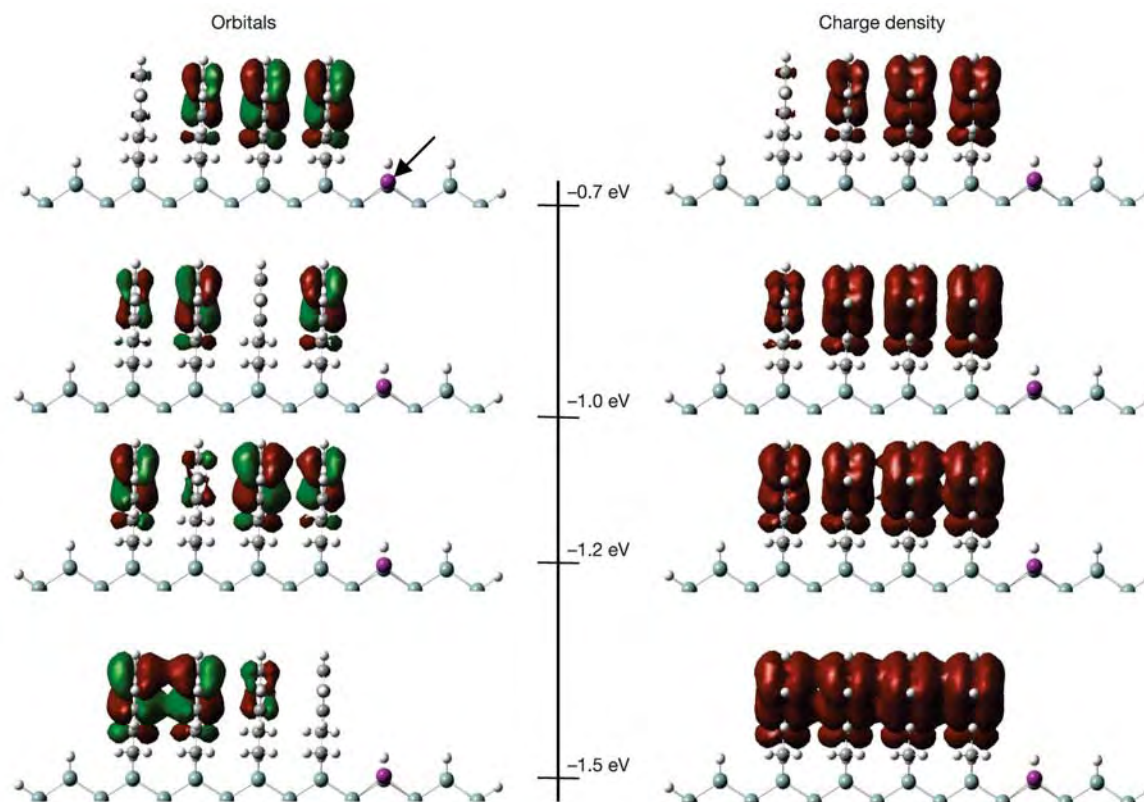


Figure 4 | Orbitals and charge densities near a dangling bond. Left, representative orbitals showing that the highest-energy molecular state is localized near the negative dangling bond (indicated by the purple sphere and arrow), while molecular states deeper in the occupied manifold are localized farther from the negative dangling bond (top to bottom). Right, charge-density surfaces of molecular states as a function of energy. Top is the charge density of the highest molecular state. Each subsequent surface

represents the sums of charge densities of molecular states from the top of the valence band to the indicated energy. These surfaces demonstrate that the slope effect appears at smaller-magnitude scan biases and disappears (images become flat) at higher-magnitude scan biases in agreement with the STM measurements. For clarity, a row of silicon dimers has been removed from the model.

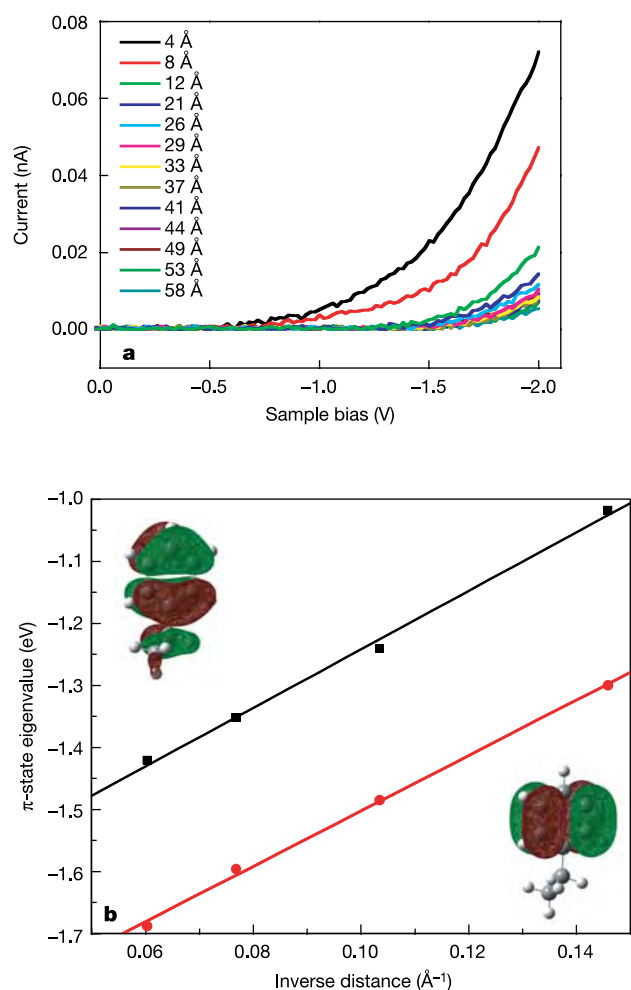


Figure 5 | Change in electronic properties with distance from the dangling bond. **a**, Current–voltage characteristics of styrene molecules at varying distances from the negatively charged dangling bond. The black spectrum (acquired closest to the dangling bond) results from a greater effective sample bias and displays a lower onset voltage and greater overall current than curves acquired farther along the molecular line. As a result of the negatively charged dangling bond's electrostatic field, the tunnel current varies by a factor of ~ 120 across the molecular line at -1.4 V for the sample. **b**, Change in the calculated π_x - and π_y -state energies (see models showing different orbital phases by colour, inset) in a single substrate-bound styrene molecule as a function of the inverse distance to the dangling bond. The linear variation shown in the plot illustrates that the orbitals are Stark-shifted by the field emanating from the charged dangling bond.

We have shown that on silicon, two contacts at distinct potentials can be made to a molecule via the dangling bond and the substrate. This has allowed us to expand the two-terminal STM experiment to gain insight into the response of a molecule to an eventual third terminal while also demonstrating the feasibility of chemical gating and pointing the way to exacting studies of field-regulated processes. It appears possible that the mechanism described here may act in a variety of environments and processes; for example, electron transport through biological molecules that intermingle with ionic species may be modulated in this way.

Received 25 January; accepted 17 March 2005.

1. Aviram, A. & Ratner, M. A. Molecular rectifiers. *Chem. Phys. Lett.* **29**, 277–283 (1974).

2. Kubatkin, S. *et al.* Single-electron transistor of a single organic molecule with access to several redox states. *Nature* **425**, 698–701 (2003).
3. Park, J. *et al.* Coulomb blockade and the Kondo effect in single-atom transistors. *Nature* **417**, 722–725 (2002).
4. Kaun, C.-C., Guo, H., Grütter, P. & Lennox, R. B. Momentum filtering effect in molecular wires. *Phys. Rev. B* **70**, 195309 (2004).
5. Yang, Z., Chshiev, M., Zwolak, M. & Di Ventra, M. Role of heating and current-induced forces in the stability of atomic wires. *Phys. Rev. B* **71**, 041402(R) (2005).
6. Damle, P., Rakshit, T., Paulsson, M. & Datta, S. Current–voltage characteristics of molecular conductors: two versus three terminal. *IEEE Trans. Nanotech.* **1**, 145–153 (2002).
7. Emberly, E. G. & Kirczenow, G. The smallest molecular switch. *Phys. Rev. Lett.* **91**, 188301 (2003).
8. Reed, M. A., Zhou, C., Muller, C. J., Burgin, T. P. & Tour, J. M. Conductance of a molecular junction. *Science* **278**, 252–254 (1997).
9. Cui, X. D. *et al.* Reproducible measurement of single-molecule conductivity. *Science* **294**, 571–574 (2001).
10. Selzer, Y. *et al.* Effect of local environment on molecular conduction: Isolated molecule versus self-assembled monolayer. *Nano Lett.* **5**, 61–65 (2005).
11. Wold, D. J., Haag, R., Rampi, M. A. & Frisbee, C. D. Distance dependence of electron tunneling through self-assembled monolayers measured by conducting probe atomic force microscopy: Unsaturated versus saturated molecular junctions. *J. Phys. Chem. B* **106**, 2813–2816 (2002).
12. Joachim, C. & Gimzewski, J. K. An electrochemical amplifier using a single molecule. *Chem. Phys. Lett.* **265**, 353–357 (1997).
13. Nazin, G. V., Qiu, X. H. & Ho, W. Visualization and spectroscopy of a metal-molecule-metal bridge. *Science* **302**, 77–81 (2003).
14. Moresco, F. *et al.* Probing the different stages in contacting a single molecular wire. *Phys. Rev. Lett.* **91**, 036601 (2003).
15. Grill, L. *et al.* Controlled manipulation of a single molecular wire along a copper atomic nanostructure. *Phys. Rev. B* **69**, 035416 (2004).
16. Mayne, A. J. *et al.* Chemisorbed bistable molecule: Biphenyl on Si(100)-2 $\times$ 1. *Phys. Rev. B* **69**, 045409 (2004).
17. Repp, J., Meyer, G., Stojković, S. M., Gourdon, A. & Joachim, C. Molecules on insulating films: Scanning-tunneling microscopy imaging of individual molecular orbitals. *Phys. Rev. Lett.* **94**, 026803 (2005).
18. Lopinski, G. P., Wayner, D. D. M. & Wolkow, R. A. Self-directed growth of molecular nanostructures on silicon. *Nature* **406**, 48–51 (2000).
19. DiLabio, G. A., Piva, P. G., Kruse, P. & Wolkow, R. A. Dispersion interactions enable the self-directed growth of linear alkane nanostructures covalently bound to silicon. *J. Am. Chem. Soc.* **126**, 16048–16050 (2004).
20. Sze, S. M. *Physics of Semiconductor Devices* Ch. 1 (Wiley-Interscience, New York, 1981).
21. Bardeen, J. Surface states and rectification at a metal-semiconductor interface. *Phys. Rev.* **71**, 717–727 (1947).
22. Feenstra, R. M., Meyer, G. & Rieder, K. H. Transport limitations in tunneling spectroscopy of Ge(111)c(2 $\times$ 8) surfaces. *Phys. Rev. B* **69**, 081309(R) (2004).
23. Pitters, J. L. & Wolkow, R. A. Protection-deprotection chemistry to control styrene self-directed line growth on hydrogen-terminated Si(100). *J. Am. Chem. Soc.* **127**, 48–49 (2005).
24. Dewar, M. J. S., Zebisch, E. G., Healy, E. F. & Stewart, J. J. P. AM1: A new general purpose quantum mechanical molecular model. *J. Am. Chem. Soc.* **107**, 3902–3909 (1985).
25. Boese, A. D. & Handy, N. C. A new parameterization of exchange-correlation generalized gradient approximation functionals. *J. Chem. Phys.* **114**, 5497–5503 (2001).
26. Stevens, W., Basch, H. & Krauss, J. Compact effective potentials and efficient shared-exponent basis sets for the first- and second-row atoms. *J. Chem. Phys.* **81**, 6026–6033 (1984).
27. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865–3868 (1996).
28. Frisch, M. J., *et al.* *Gaussian 03 Revision C.02* (Gaussian, Inc., Wallingford, Connecticut, 2004).
29. Tersoff, J. & Hamann, D. R. Theory and application for the scanning tunneling microscope. *Phys. Rev. Lett.* **50**, 1998–2001 (1983).
30. Herzberg, G. *Atomic Spectra and Atomic Structure* 2nd edn, Ch. II, 114 (Dover, New York, 1944).

Acknowledgements We have benefited from discussions with G. Kirczenow, G. Lopinski, S. Datta, H. Guo and R. Feenstra and from the technical expertise of M. Cloutier and D. Moffatt. Funding has been provided by iCORE, the NRC, the NSERC, CFI, the University of Alberta and CIAR. We are grateful for access to WestGrid and the Center of Excellence in Integrated Nanotools (University of Alberta) computational facilities.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to R.A.W. (rwolkow@ualberta.ca).

LETTERS

Arctic freshwater forcing of the Younger Dryas cold reversal

Lev Tarasov¹ & W.R. Peltier¹

The last deglaciation was abruptly interrupted by a millennial-scale reversal to glacial conditions¹, the Younger Dryas cold event. This cold interval has been connected to a decrease in the rate of North Atlantic Deep Water formation and to a resulting weakening of the meridional overturning circulation^{2–4} owing to surface water freshening. In contrast, an earlier input of fresh water (meltwater pulse 1a), whose origin is disputed^{5,6}, apparently did not lead to a reduction of the meridional overturning circulation⁴. Here we analyse an ensemble of simulations of the drainage chronology of the North American ice sheet in order to identify the geographical release points of freshwater forcing during deglaciation. According to the simulations with our calibrated glacial systems model, the North American ice sheet contributed about half the fresh water of meltwater pulse 1a. During the onset of the Younger Dryas, we find that the largest combined meltwater/iceberg discharge was directed into the Arctic Ocean. Given that the only drainage outlet from the Arctic Ocean was via the Fram Strait into the Greenland–Iceland–Norwegian seas⁷, where North Atlantic Deep Water is formed today, we hypothesize that it was this Arctic freshwater flux that triggered the Younger Dryas cold reversal.

Among the various mechanisms of climate change, those that are the most difficult to constrain and that may be the most severe are those associated with fast nonlinear processes. Abrupt and sustained changes in the thermohaline circulation (THC) have been implicated in past events of this kind, such as the Dansgaard–Oeschger oscillations that were a recurrent characteristic of marine isotope stage 3¹. It has also been suggested that similarly rapid changes in the meridional overturning circulation (MOC) could occur in response to global warming⁸. However, the actual sensitivity of North Atlantic Deep Water formation and MOC to freshwater forcing remains poorly understood. As the most recent strong millennial-scale response to a variation in the MOC, the Younger Dryas (YD) event offers a basis for a clear test of the sensitivity of North Atlantic Deep Water formation to freshwater fluxes. In order to perform this test, however, a detailed deglacial chronology of the runoff of fresh water from the continents is required. Here we analyse the largest possible contribution, that from the disintegration of the North American ice sheet (NAIS) complex.

A significant challenge to the hypothesis that it was extreme meltwater forcing that triggered the YD concerns our understanding of meltwater pulse 1a (mwp-1a). This event produced a rise of approximately 20 m in eustatic sea level over an interval of 500 yr (see below), during which no significant decrease in the MOC has been inferred⁴. $\delta^{18}\text{O}$ records from the Gulf of Mexico indicate that a large contribution to mwp-1a entered the Gulf (via the Mississippi River outlet, Fig. 1), and it has been suggested on the basis of models that such a freshwater flux would thereafter be advected into the North Atlantic by the Gulf Stream, with a resultant significant diminishing of the MOC⁹. The eustatic sea level record during the

onset of the YD, on the other hand, lacks a discernible meltwater pulse (Fig. 2). Reconstructed sea surface salinities for the Gulf of St Lawrence¹⁰ have also refuted the presence of a surface meltwater plume in this region during the period of YD onset, contradicting a previous hypothesis¹¹. Two fundamental fluid dynamical issues (see Supplementary Information) concerning the hyperpycnal behaviour of sediment-laden riverine outflow into the oceans¹² and the strong baroclinic instability of the Gulf Stream also make it unlikely that discharge of melt water into the Atlantic Ocean or the Gulf of Mexico could produce a low-salinity surface plume that was advected intact to the sites of North Atlantic Deep Water formation.

In order to significantly influence the MOC, freshwater forcing must be applied directly onto the region of North Atlantic Deep Water formation—as apparently occurred during the Heinrich events, when the surface freshening was associated with the melting of icebergs expelled into the Atlantic from the NAIS by calving through the Hudson Strait¹³. One might expect to achieve the same effect by means of freshwater delivery, especially in the form of pack ice, through the Fram Strait into the Greenland–Iceland–Norwegian (GIN) seas (Fig. 1), as has been previously hypothesized for the Preboreal Oscillation¹⁴. Indeed, four planktonic $\delta^{18}\text{O}$ data points from three sedimentary cores (PS2837, PS2887, PS1230) in the western Fram Strait^{15,16} collectively appear to indicate the presence of such an event between 10.5 and 11.2 ^{14}C kyr ago (bracketing YD onset). Our analyses demonstrate this alternative mechanism to be the preferred candidate for the cause of the YD.

In order to reconstruct a regional deglacial drainage chronology for North America that (to our knowledge, for the first time) includes an objective (though incomplete) measure of uncertainty, we use a best-fit 77 member sub-ensemble from a 5,000 member ensemble of glacial systems model (GSM) analyses¹⁷, calibrated against an extensive set of relative-sea-level and geodetic observations using a bayesian methodology. To further reduce uncertainties, the model is forced to conform to a newly developed high-resolution margin chronology derived from ^{14}C dated geological and geomorphological observations^{7,18}.

In comparing the computed deglacial contribution of the NAIS to the palaeorecord of eustatic sea level change (Fig. 2), four key points emerge. First, the GSM explains approximately half of the 20 m eustatic sea level rise associated with the mwp-1a event. Second, there is no similarly intense meltwater pulse predicted during the onset of the YD. Third, the model produces no significant contribution to mwp-1b. Fourth, as is demonstrated below, although discharge into the Gulf of Mexico and the Atlantic Ocean are the dominant contributions of the NAIS to mwp-1a, there is also substantial discharge into the Arctic Ocean (and the Pacific Ocean, though not shown here). Given the inferred collapse of the Barents Sea ice sheet during this interval¹⁹ and the stronger response of the Eurasian ice sheet to climate forcing²⁰, it follows that a substantial fraction of the remaining contribution to mwp-1a is due to Eurasian sources.

¹Department of Physics, University of Toronto, Toronto, Ontario, Canada M5S 1A7.

The lack of a direct connection between the eustatic sea level record and climate response during the period of YD onset implies that the most important factor in determining the effect of the freshwater forcing upon the MOC is not the total amount of melt water that enters the ocean basins but rather its regional distribution. Considering first the discharge into the Gulf of Mexico (Fig. 3a), the largest deglacial meltwater pulse (1.35 ± 0.4 dSv 100 yr mean, 1.65 ± 0.25 dSv after eliminating simulations with apparently unsupported Eastern routing of Lake Agassiz drainage; $1 \text{ dSv} = 10^5 \text{ m}^3 \text{ s}^{-1}$) is predicted to have occurred during the mwp-1a event, as previously inferred on the basis of the Orca basin $\delta^{18}\text{O}$ record¹¹. Southern meltwater discharge terminates by -12.9 kyr (that is, 12.9 kyr before present).

The largest total discharge into the Atlantic Ocean through a mid-latitude outlet (that is, not including the Labrador Sea) also occurs during the mwp-1a interval with a 1σ range of 1.0–1.9 dSv (Fig. 3b). Discharge into the mid-Atlantic is split between the Hudson River outlet (0.5–1.4 dSv) and the Gulf of St Lawrence (0.5 dSv). During the YD interval, significant discharge occurs only via the Gulf of St Lawrence, and has a 1σ range of 0.1–0.4 dSv (meltwater and iceberg flux only, up to 0.7 dSv including precipitation over ice-free land). Mean North American discharge into the Labrador Sea during the YD interval is, except for a peak of 0.2 dSv at -12.8 kyr, below 0.1 dSv until an imposed ice reduction representing an assumed Heinrich event H0 subsequent to -12.0 kyr.

Peak (100 yr weighted mean) ensemble discharge into the Arctic Ocean is 0.7 dSv during the mwp-1a interval (Fig. 3c). Most importantly, mean discharge first surpasses 1 dSv near -12.9 kyr and peaks at -12.8 kyr with a 1σ range of 1.2–2.2 dSv. This peak discharge into the Arctic Ocean is more than twice the sum total of all Atlantic discharge from the NAIS (including the Gulf of Mexico and the Labrador Sea) during the YD onset period. To place this in perspective, the present day outflow of the Mackenzie River is approximately 0.11 dSv, while the total present-day Arctic Ocean freshwater outflow through the Fram Strait is approximately 1.1 dSv (ref. 21). Given that our analyses ignore both Eurasian inputs into the Arctic Ocean (at

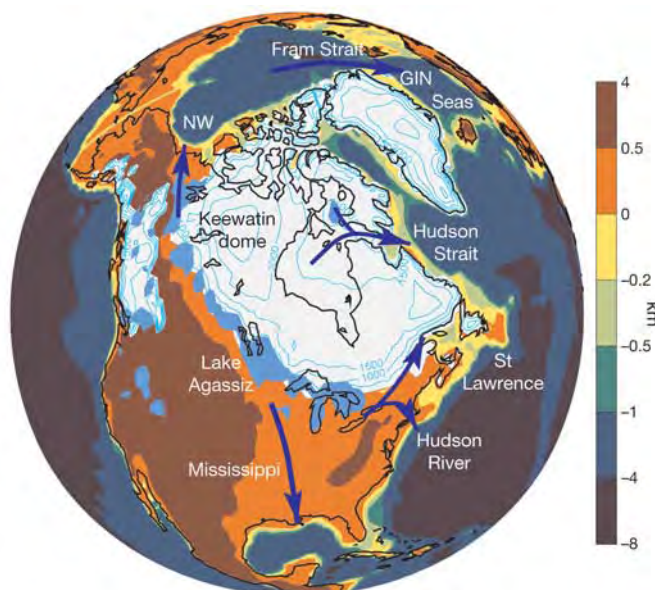


Figure 1 | Major deglacial drainage outlets for North America, along with approximate positions of proglacial Lake Agassiz and Keewatin dome just before the onset of the YD. Modern North Atlantic Deep Water formation primarily occurs in the GIN seas region. Northwest (NW) drainage of the ice complex is via the Mackenzie River basin into the Beaufort Sea and subsequently into the Canadian basin of the Arctic Ocean. Surface elevation is indicated by the colour scale.

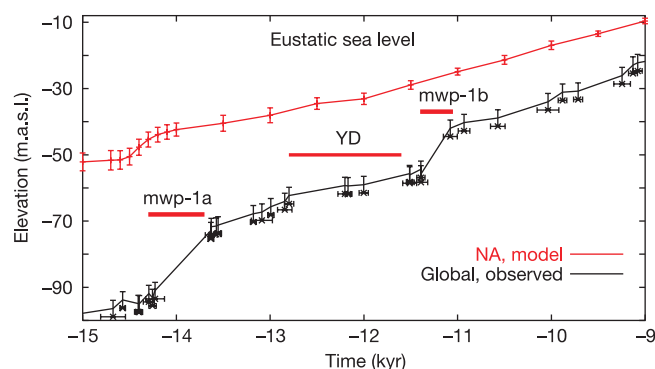


Figure 2 | Eustatic sea-level chronologies. Black curve, the observed chronology as inferred from the U/Th-dated Barbados *Acropora palmata* coral record²⁹. Red curve, the computed North American contribution ('NA, model') to eustatic sea level rise as delivered by a 78 member best-fit sub-ensemble with 1σ confidence intervals as determined by fit to the data set employed to constrain the model. Not included are uncertainties associated with the margin chronology and the limited ensemble phase space. m.a.s.l., metres above sea level. The time intervals for the mwp-1a, YD, and mwp-1b events are delineated by the horizontal red bars.

present 1.0 dSv; ref. 21) and net precipitation over the Arctic Ocean (at present 0.3 dSv; ref. 21), the 1–2 dSv increase of North American freshwater flux into the Arctic during YD onset is highly significant. Furthermore, the actual discharge would have had significant higher frequency variability (and therefore higher peak values) than is evident with the 100 yr timesteps of the drainage calculations.

Previous analyses using largely unconstrained deglacial chronologies lacking a Keewatin ice dome^{22,23} have inferred substantial discharge into the Arctic Ocean only after termination of the YD. The timing of this intense period of freshwater forcing in our reconstruction is relatively insensitive to the uncertainty in deglacial climate chronology, in that use of the GISP II $\delta^{18}\text{O}$ record for the climate forcing chronology (Fig. 3d) does not modify the timing. Rather, the timing of this event is fixed by the margin chronology and therefore subject primarily to its uncertainties (as detailed in Supplementary Discussion). The continuous high-level discharge over the whole YD interval with the inclusion of precipitation over ice-free land ('Upper bound' in Fig. 3b) may also have played a critical role in sustaining MOC reduction for a millennium.

The underlying source of this strong discharge into the Arctic Ocean is the large Keewatin ice dome (Fig. 1), whose existence at the Last Glacial Maximum was recently confirmed through analyses of space geodetic and absolute gravity constraints^{24,17}. The strength of local sourcing is evident in that even with the removal of all runs that have northwest drainage of Lake Agassiz at -12.8 kyr, ensemble discharge into the Arctic Ocean still dominates, with a 1σ range of 1.1–1.5 dSv. Our ensemble-based analyses do however indicate northwest drainage of Lake Agassiz during much or all of the YD, contrary to the eastward drainage that has until recently generally been assumed²⁵ but is now in question²⁶. The -12.8 kyr Arctic meltwater (and iceberg) flux has contributions from both the reduction of the volume of this ice dome as well as from the expansion of the drainage basin due to the isostatic depression induced by this surface load. The magnitude of this primary dome of the Last Glacial Maximum NAIS is determined by two significant constraints in the calibrated model. The recently refined observation of the rate ($6.5 \pm 1.5 \text{ mm yr}^{-1}$) of present day uplift of the surface of the solid Earth based on very-long-baseline interferometry and GPS (Global Positioning System) measurements at Yellowknife (D. F. Argus, personal communication) provides a strong regional constraint. Furthermore, global ice volume constraints together with regional limits on the amount of ice that could have existed on other

continents (based on relative sea level and glaciological analyses along with inferred ice margin chronologies) require a substantial ice load over North America, which on the basis of relative sea level observations implies large ice volume over the Keewatin region. It is also noteworthy that a large Keewatin dome was previously inferred on the basis of glacial geomorphology²⁷.

Detailed understanding of the physical process by which the freshwater flux into the Arctic Ocean affects the Atlantic MOC will require further investigation. Given the strong stable stratification that is characteristic of Arctic waters²¹, it may be that a significant surface meltwater plume was simply advected through the Fram Strait directly into the GIN seas. However, the enhanced freshwater flux would also have increased sea ice formation in the Arctic, with a

resultant enhanced flux of pack ice into the GIN seas. Given that ice transport through the Fram Strait at present accounts for three-quarters of the mean annual freshwater export from the Arctic Ocean through the Fram Strait (0.9 dSv; ref. 21), with mean monthly discharge in the winter months (probably a better analogue for YD era annual conditions) at times greater than 2.4 dSv (ref. 28), we believe that this second mechanism of freshwater transport in combination with some iceberg flux played a dominant role during YD onset. Further testing of this 'Arctic trigger' hypothesis will require improved observational constraints on the deglacial chronology of the Keewatin ice dome, and detailed data from marine sedimentary cores from the Arctic basin (especially the Beaufort Sea region), along with numerical experiments to examine directly the response of coupled atmosphere-ocean general circulation models to the deglacial drainage chronology.

METHODS

GSM. The University of Toronto Glacial Systems Model (GSM) used for the analyses presented here incorporates a three-dimensional thermo-mechanically coupled ice-sheet model, bed-thermal model, sub-glacial till-deformation model, temperature-dependent positive degree-day mass-balance model with a physical refreezing parameterization, spherically symmetric visco-elastic isotropic response model, and a fast surface drainage solver. For the results presented here, the GSM was run at 1.0° longitude by 0.5° latitude grid resolution. A complete description of the GSM is provided elsewhere (ref. 17 and references therein), and only a brief summary is provided here.

The ice-sheet component of the GSM is based upon the standard Glen flow ice rheology and shallow ice approximation. Coulomb-plastic till deformation is assumed to occur when the basal temperature approaches the pressure-melting point and adequate sediment is available. The isostatic response model employs the VM2 radial mantle viscosity structure, the PREM radial elasticity model, and a 100-km-thick surface lithosphere (with infinite viscosity). A gravitationally self-consistent relative sea level solver⁶ is applied in post-processing to compare model predictions to observations. The climate forcing used to drive the GSM is derived from the GRIP $\delta^{18}\text{O}$ record in combination with reconstructed isotopic sensitivity parameters to define a glacial index that is used to linearly interpolate between a glacial climate state derived from an ensemble of PMIP (Paleo Model Intercomparison Project, <http://www-lsce.cea.fr/pmip/>) general circulation model reconstructions of Last Glacial Maximum climate and a present day reanalysis-based climatology.

Model calibration. 22 ensemble parameters representing uncertainties in climate forcing, ice calving, and fast flow physics (including forced ice reduction during Heinrich events 1 and 0) are varied in a bayesian calibration (R. Neal, W.R.P. and L.T., manuscript in preparation) of the GSM against a large set (over 5,540 data points) of relative-sea-level and geodetic observations. The bayesian calibration employs a multilayer perceptron neural network simulator of the GSM to extensively probe the model phase space. The posterior distribution for a parameter set given the constraint data set is proportional to the product of the prior probability distribution of the parameters and the probability of the observational constraint data given the parameters. Trial parameter sets are extracted by means of Markov Chain Monte Carlo (MCMC) sampling from this posterior distribution. These parameter sets are then applied to the full GSM. Subsequent results are employed to further train the neural network (using bayesian methods) for further iterative calibration. Each ensemble run covers a whole glacial cycle starting at -122 kyr (that is, the Eemian interglacial). The limited phase space of the GSM in combination with uncertainties in the applied margin chronology^{7,18} and the regionally (and temporally) limited coverage of the constraint data set constitute the largest uncertainties that cannot be accounted for in the calibration procedure.

Surface drainage. The surface drainage solver diagnostically computes down-slope drainage along with surface water storage using a two-stage depression-fill algorithm, and is run at the same spatial resolution as the rest of the GSM. In essence, the algorithm is quite simple: diagnostically let meltwater (averaged over 100 yr) flow down the contemporaneous surface slope at the end of the 100 yr diagnostic time-step, filling depressions (subject to available melt water), until it either enters a depression for which not enough melt water is available to overflow the depression, or until melt water reaches the deep ocean (defined as regions with present-day bathymetry deeper than 600 m). The meltwater discharge into each ocean basin (Pacific, Arctic, Labrador Sea, St Lawrence (Atlantic, south of Newfoundland and excluding Hudson River basin), Hudson River basin, Caribbean (Mississippi)) then follows from the summation of the meltwater fluxes into the deep-water sector of each basin. To provide a sense of

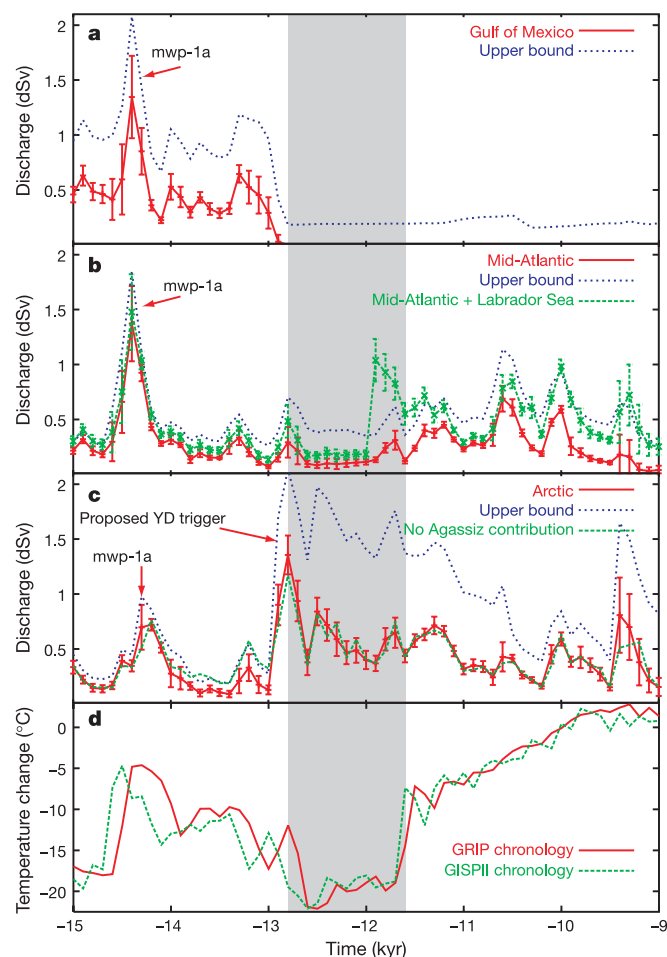


Figure 3 | Computed regional drainage chronologies and the inferred regional temperature change chronology. **a–c,** Computed regional drainage chronologies for the Gulf of Mexico (**a**), the Atlantic (**b**) and the Arctic (**c**). **d,** The inferred regional temperature change chronology from Central Greenland, from a calibrated glaciological model³⁰. 'Mid-Atlantic' (**b**) is all discharge from Newfoundland to Georgia. 'No Agassiz contribution' (**c**) is the mean Arctic discharge for ensemble runs that have eastern drainage of Lake Agassiz at -12.8 kyr. The large scatter in Gulf of Mexico (**a**) and Atlantic (**b**) discharge at -14.4 kyr is due to variations in the routing of Lake Agassiz drainage between eastern and southern outlets. The grey bar denotes the YD interval. Surface drainage is computed every 100 yr using mean meltwater (from the ice-sheet only) and iceberg fluxes over the 100 yr interval and the instantaneous surface topography. The 1 σ confidence intervals shown are as per Fig. 2. 'Upper bound' denotes the 1 σ upper bound with the additional inclusion of precipitation over ice-free land in the discharge calculation. Interpretation of these latter results requires recognition of the large uncertainties in estimated deglacial precipitation over ice-free land.

the range of drainage basin configurations for a single timeslice, computed –12.8 kyr drainage basins for three good-fit runs are shown in the Supplementary Figures. The algorithm assumes zero water depth across controlling sills. Meltwater and iceberg discharge are lumped together in the surface drainage determination. The drainage topography used in the GSM is derived from the high resolution hydrologically correct HYDRO1K digital elevation map (<http://edcdaac.usgs.gov/gtopo30/hydro/namerica.asp>) with local modifications based on sub-grid manual checks of critical drainage choke-points. A few critical choke-point elevations are taken to be time-dependent, to account for either sub-grid movement of the ice margin across the grid-cell (determined using linear interpolation between ice margin chronology time-slices across a much higher resolution version of the drainage topography) or erosional changes inferred on the basis of regional strandline data. Changes in surface water loading (excluding geoidal perturbations in the coupled model) are also accounted for in computing the isostatic adjustment of the solid Earth. The combined solver and drainage topography have been verified against a coarse-grained version of the level 1 drainage basins of the HYDRO1k data set.

Received 5 October 2004; accepted 1 April 2005.

- Dansgaard, W. *et al.* Evidence for general instability of past climate from a 250 kyr ice-core record. *Nature* **264**, 218–220 (1993).
- Keigwin, L. D., Jones, J. A., Lehman, S. J. & Boyle, E. A. Deglacial meltwater discharge, North-Atlantic deep circulation, and abrupt climate change. *J. Geophys. Res.* **96**, 16811–16826 (1991).
- Muscheler, R., Beer, J., Wagner, G. & Finkel, R. Changes in deep-water formation during the Younger Dryas event inferred from ^{10}Be and ^{14}C records. *Nature* **408**, 567–570 (2000).
- McManus, J. F., Francois, R., Gherardi, J.-M., Keigwin, L. D. & Brown-Leger, S. Collapse and rapid resumption of Atlantic meridional circulation linked to deglacial climate changes. *Nature* **428**, 834–837 (2004).
- Clark, P. *et al.* Origin of the first global meltwater pulse following the last glacial maximum. *Paleoceanography* **11**, 563–577 (1996).
- Peltier, W. R. On the hemispheric origins of meltwater pulse 1-a. *Quat. Sci. Rev.* (in the press).
- Dyke, A. S. in *Quaternary Glaciations—Extent and Chronology*, Part II, Vol. 2b (eds Ehlers, J. & Gibbard, P. L.) 373–424 (Elsevier Science and Technology Books, Amsterdam, 2004).
- Stocker, T. F. & Schmittner, A. Influence of carbon dioxide emission rates on the stability of the thermohaline circulation. *Nature* **388**, 862–865 (1997).
- Manabe, S. & Stouffer, R. J. Coupled ocean-atmosphere model response to freshwater input: comparison to Younger Dryas event. *Paleoceanography* **12**, 321–336 (1997).
- de Vernal, A., Hillaire-Marcel, C. & Bilodeau, G. Reduced meltwater outflow from the Laurentide ice margin during the Younger Dryas. *Nature* **381**, 774–777 (1996).
- Broecker, W. S. *et al.* Routing of meltwater from the Laurentide Ice Sheet during the Younger Dryas cold episode. *Nature* **341**, 318–321 (1989).
- Parsons, J. D., Bush, J. W. & Syvitski, J. P. Hyperpycnal plume formation from riverine outflows with small sediment concentrations. *Sedimentology* **48**, 465–478 (2001).
- Hemming, S. R. Massive late Pleistocene detritus layers of the North Atlantic and their global climate imprint. *Rev. Geophys.* **42**, doi:10.1029/2003RG000128 (2004).
- Fisher, T. G., Smith, D. G. & Andrews, J. T. Preboreal oscillation caused by a glacial Lake Agassiz flood. *Quat. Sci. Rev.* **21**, 873–878 (2002).
- Bauch, H. A. *et al.* A multiproxy reconstruction of the evolution of deep and surface waters in the subarctic Nordic seas over the last 30,000 yr. *Quat. Sci. Rev.* **20**, 659–678 (2001).
- Norgaard-Pedersen, N. *et al.* Arctic Ocean during the Last Glacial Maximum: Atlantic and polar domains of surface water mass distribution and ice cover. *Paleoc.* **18**, doi:10.1029/2002PA000781 (2003).
- Tarasov, L. & Peltier, W. R. A geophysically constrained large ensemble analysis of the deglacial history of the North American ice sheet complex. *Quat. Sci. Rev.* **23**, 359–388 (2004).
- Dyke, A. S., Moore, A. & Robertson, L. *Deglaciation of North America* (Tech. Rep. Open File 1574, Geological Survey of Canada, Ottawa, 2003).
- Svendsen, J. I. *et al.* Late Quaternary ice sheet history of northern Eurasia. *Quat. Sci. Rev.* **23**, 1229–1271 (2004).
- Tarasov, L. & Peltier, W. R. Terminating the 100 kyr ice age cycle. *J. Geophys. Res.* **102**, 21665–21693 (1997).
- Aagaard, K. & Carmack, E. C. The role of sea ice and other fresh water in the Arctic circulation. *J. Geophys. Res.* **94**, 14485–14498 (1989).
- Licciardi, J. M., Teller, J. T. & Clark, P. U. in *Mechanisms of Global Climate Change at Millennial Time Scales* (eds Clark, P. U., Webb, R. S. & Keigwin, L. D.) 177–201 (AGU Geophysical Monographs Vol. 112, American Geophysical Union, Washington DC, 1999).
- Marshall, S. J. & Clarke, G. K. C. Modeling North American freshwater runoff through the last glacial cycle. *Quat. Res.* **52**, 300–315 (1999).
- Peltier, W. R. Global glacial isostatic adjustment: Paleo-geodetic and space-geodetic tests of the ICE-4G (VM2) model. *J. Quat. Sci.* **17**, 491–510 (2002).
- Teller, J. T. & Leverington, D. W. Glacial Lake Agassiz: a 5000 yr history of change and its relationship to the $\delta^{18}\text{O}$ record of Greenland. *Geol. Soc. Am. Bull.* **116**, 729–742 (2004).
- Teller, J. T., Boyd, M., Yang, Z., Kor, P. S. G. & Fard, A. M. Alternative routing of Lake Agassiz overflow during the Younger Dryas: New dates, paleotopography, and a reevaluation. *Quat. Sci. Rev.* (in the press).
- Dyke, A. S. & Prest, V. K. Late Wisconsinan and Holocene history of the Laurentide ice sheet. *Géogr. Phys. Quat.* **41**, 237–264 (1987).
- Vinje, T., Nordlund, N. & Kvambekk, A. Monitoring ice thickness in Fram Strait. *J. Geophys. Res.* **103**, 10437–10449 (1998).
- Fairbanks, R. G. A 17,000-year glacio-eustatic sea level record: influence of glacial melting rates on the Younger Dryas event and deep-ocean circulation. *Nature* **342**, 637–641 (1989).
- Tarasov, L. & Peltier, W. R. Greenland glacial history, borehole constraints and Eemian extent. *J. Geophys. Res.* **108**, 2124–2143 (2003).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This paper is a contribution to the Polar Climate Stability Research Network, which is funded by the Canadian Foundation for Climate and Atmospheric Sciences and a consortium of Canadian universities. We thank W. Broecker, A. Dyke, T. Fisher, C. Hillaire-Marcel and J. Teller for discussions.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to L.T. (lev@atmos.physics.utoronto.ca).

LETTERS

Seismological constraints on a possible plume root at the core-mantle boundary

Sebastian Rost¹, Edward J. Garnero¹, Quentin Williams² & Michael Manga³

Recent seismological discoveries have indicated that the Earth's core-mantle boundary is far more complex than a simple boundary between the molten outer core and the silicate mantle. Instead, its structural complexities probably rival those of the Earth's crust¹. Some regions of the lowermost mantle have been observed to have seismic wave speed reductions of at least 10 per cent²⁻⁷, which appear not to be global in extent⁷⁻⁹. Here we present robust evidence for an 8.5-km-thick and ~50-km-wide pocket of dense, partially molten material at the core-mantle boundary east of Australia. Array analyses of an anomalous precursor to the reflected seismic wave ScP reveal compressional and shear-wave velocity reductions of 8 and 25 per cent, respectively, and a 10 per cent increase in density of the partially molten aggregate. Seismological data are incompatible with a basal layer composed of pure melt, and thus require a mechanism to prevent downward percolation of dense melt within the layer. This may be possible by trapping of melt by cumulus crystal growth following melt drainage from an anomalously hot overlying region of the lowermost mantle. This magmatic evolution and the resulting cumulate structure seem to be associated with overlying thermal instabilities, and thus may mark a root zone of an upwelling plume.

Partial melting of lowermost mantle rock^{10,11} and chemical reactions between mantle and core material¹²⁻¹⁴ are viable candidates for causing ultralow velocity zone (ULVZ) structure at Earth's core-mantle boundary (CMB); each has distinct effects on seismic properties. Determining the likely cause(s) for ULVZ hinges on constraining ULVZ spatial extent, density and seismic velocity^{10,14,15}, and the sharpness of the boundary between ULVZ and overlying mantle. We investigate the CMB east of Australia, slightly south of New Caledonia, where P- and S-wave velocity (v_P and v_S , respectively) reductions are seen in the lower mantle^{16,17}, and complex CMB properties have been detected in isolated locations^{5,7,18}. In this study, an unprecedented high-resolution CMB sampling over a 100×250 km region is achieved using a newly assembled data set of 305 Tonga-Fiji earthquakes recorded at the small aperture (20 km) Warramunga seismic array (WRA), equipped with 20 short-period (~1 Hz dominant period) vertical-component seismometers (Fig. 1a). This level of spatial sampling distinguishes our study from previous, broader wavelength ULVZ investigations. Earthquakes were chosen on the basis of simple and similar source-time functions, and deep focal depths (>450 km) to minimize upper mantle attenuation and scattering effects and to avoid surface reflected phases. Data displaying complex source-time functions, clipping, or low signal-to-noise ratio (SNR) were omitted.

The seismic wave ScP travels from earthquake to CMB as an S wave, converts to a P wave upon reflection, and returns to the surface as a P wave. ScP waveforms from each event were beam-formed, a method that stacks data in the direction that maximizes ScP

coherency and amplitude, and significantly increases the SNR while minimizing possible influence of crustal structure beneath array stations. Figure 1b shows ScP stacks for the most densely sampled CMB region in our study area. Records between -24.60° and -25.10° in latitude show a very coherent precursor ~1.8 s before ScP. Separately summing the simple and anomalous records highlights the precursor existence (Fig. 1c). Precursor evidence vanishes less than ~20 km from the central anomalous zone; several records sampling this transition exhibit complex behaviour, suggesting multipathing.

Interaction of ScP with the top of a ULVZ generates a precursor^{4,5,19} by reflection (and conversion) from S-to-P at the top of the ULVZ, as well as an S-to-P conversion at the ULVZ surface, which continues downward as a P wave. ULVZ structure can also generate later arriving phases that are more difficult to detect and analyse, as they arrive in the ScP coda. In our data set, the dominant precursory energy is a conversion at the top of the ULVZ instead of the CMB, thus arriving sooner. A very compact region of the study area roughly 50 km in dimension is dominated by clustering of the highest quality ULVZ precursors (Fig. 2). This small dimension with well resolved sharp edges is significantly smaller than that resolvable in previous studies^{6,7}.

ScP precursor timing, amplitude and waveform were modelled with gaussian beam²⁰ synthetic seismograms (Fig. 3a). All dominant ULVZ P or S reflections, reverberations and conversions were included in the modelling, as well as a source mechanism that fits observed P waveforms. We explored a wide parameter space (Fig. 3), and constrain best-fit model properties: 8.5 ± 1 km thick, $8 \pm 2.5\%$ v_P and $25 \pm 4\%$ v_S reductions, and $10 \pm 5\%$ density increase (Fig. 3b). Models lacking a density increase poorly fit the observed waveforms; as seen in Fig. 3, ScP precursory energy is very sensitive to ULVZ density, in fact tolerating only slight offsets ($\leq 2\%$) from the best fit 10% value before waveform degradations occur. The geometric and elastic uncertainties are substantially less than those of previous ULVZ studies^{3,7,21}. We note that our finding of a 10% density anomaly is considerably less than the 18–33% that have been found for ULVZ generated by iron-rich subducted sediments²². D'' anisotropy (for example, due to the post-perovskite phase¹⁵) may split the S part of ScP. But to yield ScP with a precursor, two pulses with SV energy are required; furthermore, the anisotropic zone needs to be several hundred kilometres thick (to match the 1.8 s precursor time) and limited to a narrow corridor around the paths displaying precursors. Both of these requirements are less likely than the simpler ULVZ structure we propose here.

A diagram of the best-fitting ULVZ structure is shown in Fig. 4. The elevated density reflects the aggregate density of the liquid and solid of the partially molten assemblage; the relative density contrast between liquid and solid is unclear. Our modelling is principally

¹Department of Geological Sciences, Arizona State University, Box 871404, Tempe, Arizona 85287-1404, USA. ²Department of Earth Sciences, University of California Santa Cruz, 1156 High Street, Santa Cruz, California 95064, USA. ³Department of Earth and Planetary Science, University of California Berkeley, 307 McCone Hall, Berkeley, California 94720-4767, USA.

sensitive to average ULVZ properties; the internal zonation is not resolved. However, the upper ULVZ boundary is seismically sharp, with a probable width of less than ~ 1 km: the rapid onset is consistent with most (if not all) of the velocity decrement occurring near the ULVZ surface.

We performed slowness-backazimuth stacks, and rule out strong reflections from off-great-circle paths as a source of the precursors. We also rule out multipathing or scattering in the source region, as earthquakes that produce the ScP precursors are distributed over a 170 km depth range, and over a 90 km linear section of the Fiji-Tonga trench; a strong scatterer would be required roughly 20 km from each source over this dimension, which should produce more precursor travel time variation. Neighbouring sources should detect it as well, from a variety of distances.

Our best-fit v_S -to- v_P reduction of 25:8 is very close to the expected 3:1 for partial melt^{10,14}; an 8% v_P decrease can be explained by a 5–30% volume fraction of melt, depending on the melt geometry¹⁰. Partial melting is therefore a strong candidate for explaining our observations, and consistent with higher than average temperatures producing observed low (tomographically derived) seismic velocities in this region. A fully molten layer produces waveforms incompatible with observations. The presumably low-viscosity, dense, partially molten material clearly has not spread out along the CMB; neither

has the liquid (certainly denser than the overlying mantle material) drained downwards and segregated from its neighbouring solids.

The notable increase in density of the region implies that a simple elevation of isotherms^{7,21} is unlikely to explain fully the presence of this ULVZ micro-patch: a coupled dynamical and chemical, rather than solely thermal, explanation is required. The sharpness and high melt fraction are not consistent with simple compaction of a partially molten zone²³—and dense melt produced in an upwelling will drain downward, drainage will be most effective where the porosity (and hence permeability) is high, and a diffuse low-melt-fraction region should thus persist on the top of the layer.

One key seismological inference is the apparently uniform ULVZ property of proximity to the critical percolation threshold of liquid, if liquid is present in isolated and randomly-positioned pockets²⁴. Indeed, a principal enigma of ULVZ models is that partial melt exists without apparent segregation over broad regions of the CMB. High matrix permeability from interconnected melt requires no density difference between coexisting liquid and solids in order to maintain partial melt throughout the ULVZ (an unlikely but not impossible situation²⁵). This is not consistent with the seismic modelling.

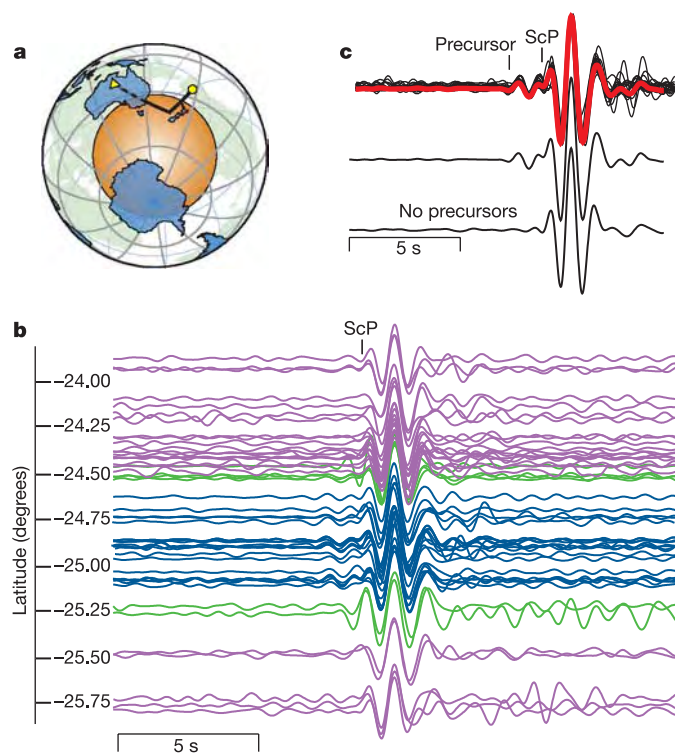


Figure 1 | Array beams of ScP and precursors. **a**, Source-receiver combination from earthquakes in the Tonga-Fiji region (circle) to the Warramunga array (triangle). **b**, Array beam traces of the original short-period array recordings aligned on the ScP arrival sorted with respect to the latitude of the ScP CMB reflection point, from -23.8° to -25.75° in a north-south profile. The epicentral distance range spanned by these data are 40.8° to 46.8° . The longitudinal extension of the sources is only a few degrees owing to the configuration of sources in the Tonga-Fiji subduction zone. The processing of the data are restricted to filtering with a second order bandpass with cut-off frequencies of 0.5 Hz and 1.4 Hz. Blue traces show a precursor arriving 1.8 s before ScP. Green traces show complicated waveforms with evidence of anomalous multipathing. Purple traces show no waveform complications. **c**, Precursory traces (blue signals in **b**) in comparison to events lacking the precursor (purple signals in **b**). Also shown are the summation traces for the precursor and non-precursor waveforms.

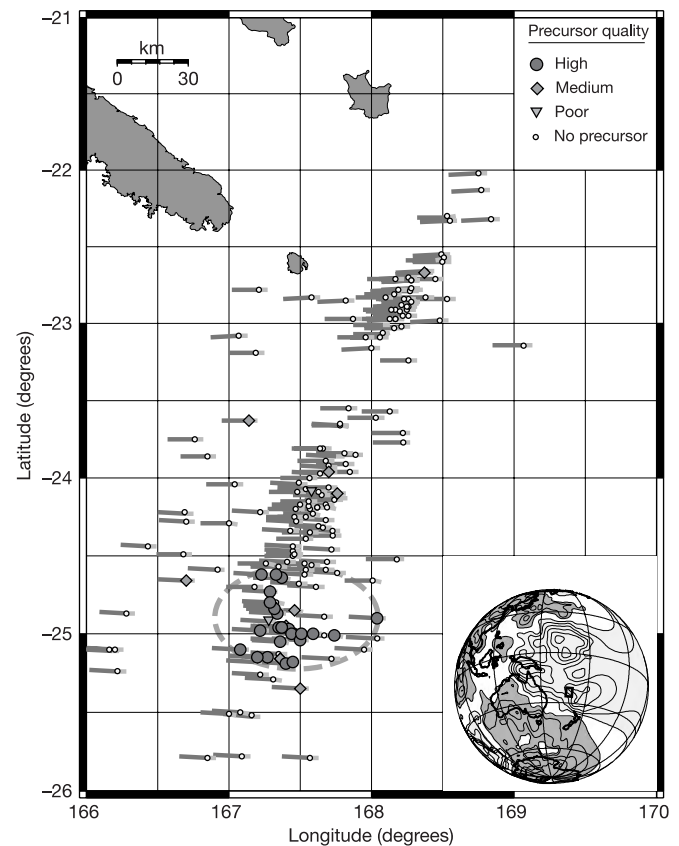


Figure 2 | ULVZ detections at ScP core-reflection locations. Clear non-detections of precursory energy to ScP are marked with small open circles. Short grey horizontal line segments denote the S- and P-wave segments (to the right and left of symbols, respectively) of ScP in an 8-km-thick ULVZ. Precursor quality is also noted, which is a measure of the precursor amplitude compared to background signal-to-noise ratio. The dashed grey line shows the region with the coherent precursor shown in Fig. 1. Inset globe displays the tomographically derived shear-wave heterogeneity²⁰ for the lowermost 200 km of the mantle, upon which the region of the larger map is denoted (small rectangle towards centre of globe). Lighter and darker shading corresponds to lower and higher relative wave speeds, respectively. Contour lines are in 0.5% intervals, and the interval $-0.5\% < v_s < +0.5\%$ is not shaded. Our study region coincides with a very-low-velocity region.

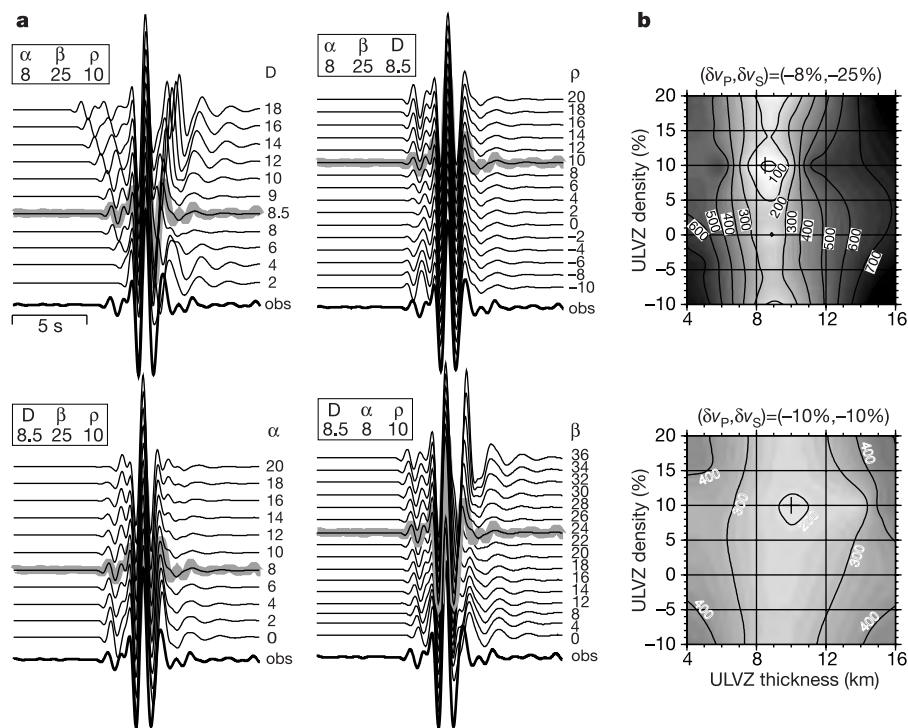


Figure 3 | Synthetic waveform modelling of ScP precursors. **a**, Synthetic seismograms for varying ULVZ thickness, v_p/v_s and density. For comparison, the ScP precursor data trace is shown at the bottom of each panel and as a light grey line beneath the best-fitting model of each model class. The best-fit model has an 8.5-km-thick ULVZ with P and S reductions of 8% and 25%, respectively, and a 10% density increase. Models with differing thickness ('D'), P and S velocity decreases (' α ' and ' β ', respectively), and density increases (' ρ ') are shown. We note that a strong density increase is necessary to fit the observed waveforms. Additional waveform modelling is presented in Supplementary Information. **b**, Goodness of fit of

synthetic predictions to data. The quality of fit is calculated by computing a residual trace by subtracting synthetic waveforms from the data in a 10 s time window around the ScP arrival, taking the envelope of the residual trace, and integrating (thus, a proxy for residual energy representing misfit). Two fixed v_p/v_s ratios (8:25 and 10:10) are shown for varying ULVZ thickness and density. Models with v_p/v_s of 1:1 show a worse fit to the data: the minimum misfit energy in the 8:25 plot is 15, compared to 188 for the 10:10 case. The detection level threshold for ULVZ thickness in this study is ~ 4 km.

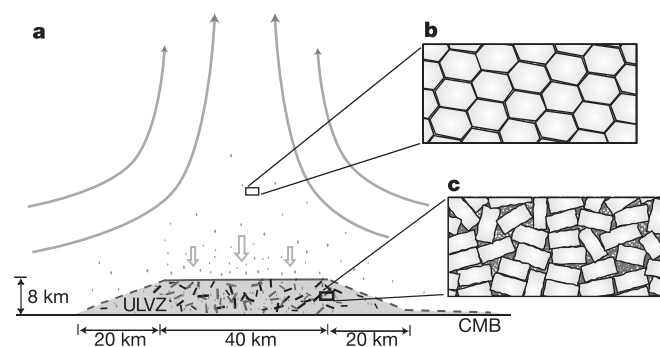


Figure 4 | Preferred model of dense partially molten ULVZ. **a**, ULVZ structural parameters of best-fitting model, in a south to north cross-section corresponding to dashed region in Fig. 2. Dense melt from above percolates downward (**b**); whether this melt chemically re-equilibrates with its neighbouring solids while descending is unknown. Melt accumulates within the ULVZ (**c**), where it is sequestered as an intercumulus liquid at concentrations just below the percolation threshold. Both intercumulus crystal growth and compaction effects are probably critical in determining the properties of this zone. The precise crystal dimensions (black dashes in ULVZ region) are uncertain, but coarsening of crystals via intercumulus growth could markedly enhance crystal sizes within the ULVZ relative to those within the overlying mantle. The amount of melt in the mantle region above the CMB probably does not exceed 0.2%, based on the size of observed seismic velocity decrements, while that within the ULVZ is near 25%.

Alternatively, melt may be trapped within the ULVZ as an intercumulus liquid: crystals within an initially partial molten system are progressively overgrown, trapping residual (and incompatible-element enriched) liquid. The elevated density may be associated with iron partitioning into the silicate liquid, and resultant enrichment of the ULVZ in iron²⁶; progressive cumulate growth is likely to generate iron-rich crystals coexisting with these liquids. The wetting behaviour (and chemistry) of such melts is ill-constrained. However, dense melts that readily form interconnected networks are likely to descend to the core: therefore, the ULVZ may only retain liquids whose wetting behaviour allows them to be trapped. The genesis of a cumulate layer is probably related to downward melt percolation and intercumulus crystal growth as melt interacts with lowermost mantle solids, compatible with the seismic observations.

The geodynamic implications of such partial melting and cumulate formation are profound; large overlying thermal anomalies are required, which in turn are probably connected to dynamic instabilities that give rise to mantle plumes. This is consistent with the general correlation between ULVZ locations, reduced lowermost mantle velocities^{7,27}, and hotspot volcanism²⁷. Furthermore, if the lowermost mantle is adiabatic and the slope of the mantle solidus is positive and larger than the adiabat, such cumulus growth and melt trapping should occur as melts partially resolidify on descent²⁸. Thus, the solid matrix of the ULVZ may contain relatively large, annealed crystals whose chemistry is modulated by precipitation of the coexisting liquid, juxtaposed with pockets of strongly incompatible element enriched melt trapped in intercumulus regions. It is possible that continual or episodic recharge (and possible discharge into the

core) of melt keeps the system near the critical percolation threshold for melt.

The seismically determined ULVZ dimensions, coupled with the well-constrained large density contrast, provide constraints on lower mantle dynamics and properties. For example, if we assume that the shape of the partial melt pocket is approximately steady over geological timescales, and that ULVZ patches form below upwellings (Fig. 4), then ULVZ thickness and radial extent can be estimated by balancing viscous stresses caused by the upwelling flow in the surrounding mantle with buoyancy forces that will cause the layer to spread (see Supplementary Information). A density difference of 10%, a thickness of 8 km and radial extent of 25 km, and a viscosity of 10^{19} Pa s imply a reasonable upwelling velocity of 0.1 m yr^{-1} .

The existence of melt in this very-low-seismic-velocity, and presumably hot, region of D'' is not unexpected. Tomographic images do not have sufficient resolution to image small-scale plume-like features, but in a recent study²⁹ broader scale (greater than ~ 200 km) plume features have been imaged, and low velocities in our study region are suggested. Geodynamic modelling of mantle flow based on tomographic models³⁰ indicates that the roots of three plumes (Tasmanid, Lord Howe and East Australia) are in the vicinity of our study area, but the CMB-to-surface extension of these features remains unclear.

The processes of melt segregation and drainage from above, coupled with chemical modification of the liquid by postcumulus crystal growth and/or interaction with the core, will each serve to generate a liquid enriched in incompatible elements. Therefore, if heat-producing elements continue to be partitioned into silicate liquid at CMB conditions, we expect this type of zone to not only be initially generated by melt drainage from a hot thermal upwelling, but also to function as a long-lived root for upwellings³¹. As a result, such localized patches of dense ULVZ melt may produce a fixed base for hot upwellings, and thus plumes: indeed, the genesis and longevity of whole mantle plumes may be intimately tied to the production of a ULVZ at their base.

Received 2 February; accepted 6 April 2005.

- Garnero, E. J. Heterogeneity of the lowermost mantle. *Annu. Rev. Earth Planet. Sci.* **28**, 509–537 (2000).
- Mori, J. & Helmberger, D. V. Localized boundary layer below the mid-Pacific velocity anomaly from a PcP precursor. *J. Geophys. Res.* **100**, 20359–20365 (1995).
- Garnero, E. J. & Helmberger, D. V. Further structural constraints and uncertainties of a thin laterally varying ultralow-velocity layer at the base of the mantle. *J. Geophys. Res.* **103**, 12495–12509 (1998).
- Reasoner, C. & Revenaugh, J. ScP constraints on ultralow-velocity zone density and gradient thickness beneath the Pacific. *J. Geophys. Res.* **105**, 28173–28182 (2000).
- Rost, S. & Revenaugh, J. Small-scale ultralow-velocity zone structure imaged by ScP. *J. Geophys. Res.* **108**(B1), 2056, doi:10.1029/2001JB001627 (2003).
- Rondenay, S. & Fischer, K. M. Constraints on localized core-mantle boundary structure from multichannel, broadband SKS coda analysis. *J. Geophys. Res.* **108**(B11), 2537, doi:10.1029/2003JB002518 (2003).
- Thorne, M. & Garnero, E. J. Inferences on ultralow-velocity zone structure from a global analysis of SPdKS waves. *J. Geophys. Res.* **109**, B08301, doi:10.1029/2004JB003010 (2004).
- Castle, J. C. & Van der Hilst, R. D. The core-mantle boundary under the Gulf of Alaska: No ULVZ for shear waves. *Earth Planet. Sci. Lett.* **176**, 311–321 (2000).
- Persh, S. E. & Vidale, J. E. Reflection properties of the core-mantle boundary from global stacks of PcP and ScP. *J. Geophys. Res.* **109**, B04309, doi:10.1029/2003JB002768 (2004).
- Williams, Q. & Garnero, E. J. Seismic evidence for partial melt at the base of Earth's mantle. *Science* **273**, 1528–1530 (1996).
- Vidale, J. E. & Hedlin, M. A. H. Evidence for partial melt at the core-mantle boundary north of Tonga from the strong scattering of seismic waves. *Nature* **391**, 682–685 (1998).
- Knittle, E. & Jeanloz, R. Earth's core-mantle boundary: results of experiments at high pressures and temperatures. *Science* **251**, 1438–1443 (1991).
- Song, X. & Ahrens, T. J. Pressure-temperature range of reactions between liquid iron in the outer core and mantle silicates. *Geophys. Res. Lett.* **21**, 153–156 (1994).
- Berryman, J. G. Seismic velocity decrement ratios for regions of partial melt in the lower mantle. *Geophys. Res. Lett.* **27**, 421–424 (2000).
- Murakami, M., Hirose, K., Kawamura, K., Sata, N. & Ohishi, Y. Post-perovskite phase transition in MgSiO_3 . *Science* **304**, 855–858 (2004).
- Karason, H. & van der Hilst, R. D. Tomographic imaging of the lowermost mantle with differential times of refracted and diffracted core phases (PKP, P_{diff}). *J. Geophys. Res.* **106**, 6569–6587 (2001).
- Grand, S. P. Mantle shear-wave tomography and the fate of subducted slabs. *Phil. Trans. R. Soc. Lond. A* **360**, 2475–2491 (2002).
- Rost, S. & Revenaugh, J. Seismic detection of rigid zones at the top of the core. *Science* **294**, 1911–1914 (2001).
- Garnero, E. J. & Vidale, J. E. ScP: a probe of ultralow velocity zones at the base of the mantle. *Geophys. Res. Lett.* **26**, 377–380 (1999).
- Cerveny, V. & Psencik, I. Gaussian beams in elastic 2-D laterally varying layered structures. *Geophys. J. Int.* **78**, 65–91 (1984).
- Garnero, E. J., Revenaugh, J., Williams, Q., Lay, T. & Kellogg, L. H. in *The Core-Mantle Boundary Region* (eds Gurnis, M., Wysession, M., Knittle, E. & Buffett, B.) 319–334 (Geodynamics series, Vol. 28, American Geophysical Union, Washington DC, 1998).
- Dobson, D. P. & Brodholt, J. P. Subducted iron formations as a source of ultralow-velocity zones at the core-mantle boundary. *Nature* **434**, 371–374 (2005).
- McKenzie, D. P. The extraction of magma from the crust and mantle. *Earth Planet. Sci. Lett.* **74**, 81–91 (1985).
- Roscoe, R. The viscosity of suspensions of rigid spheres. *J. Appl. Phys.* **3**, 267–269 (1952).
- Akins, J. A., Luo, S.-N., Asimow, P. D. & Ahrens, T. J. Shock-induced melting of MgSiO_3 perovskite and implications for melts in Earth's lowermost mantle. *Geophys. Res. Lett.* **31**(14), doi:10.1029/2004GL020237 (2004).
- Knittle, E. in *The Core-Mantle Boundary Region* (eds Gurnis, M., Wysession, M., Knittle, E. & Buffett, B.) 119–130 (Geodynamics Series, Vol. 28, American Geophysical Union, Washington DC, 1998).
- Williams, Q., Revenaugh, J. & Garnero, E. J. A correlation between ultra-low basal velocities in the mantle and hot spots. *Science* **281**, 546–549 (1998).
- Walker, D., Agee, C. & Zhang, Y. Fusion curve slope and crystal/liquid buoyancy. *J. Geophys. Res.* **93**, 313–323 (1988).
- Montelli, R. et al. Finite-frequency tomography reveals a variety of plumes in the mantle. *Science* **303**, 338–343 (2004).
- Steinberger, B. Plumes in a convecting mantle: Models and observations for individual hotspots. *J. Geophys. Res.* **105**, 11127–11152 (2000).
- Jellinek, A. M. & Manga, M. The influence of a chemical boundary layer on the fixity, spacing and lifetime of mantle plumes. *Nature* **418**, 760–763 (2002).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank R. van der Hilst and S. Grand for supplying tomographic models, S. Grand for a Futterman t^* code, and the Seismological group of MoD at Blacknest for the WRA data set. This research was supported by NSF.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to S.R. (srost@asu.edu).

LETTERS

Discovery of a short-necked sauropod dinosaur from the Late Jurassic period of Patagonia

Oliver W. M. Rauhut^{1,2}, Kristian Remes¹, Regina Fechner¹, Gerardo Cladera² & Pablo Puerta²

Sauropod dinosaurs are one of the most conspicuous groups of Mesozoic terrestrial vertebrates. They show general trends towards an overall increase in size and elongation of the neck, by means of considerable elongation of the length of individual vertebrae and a cervical vertebra count that, in some cases, increases to 19 (ref. 1). The long neck is a particular hallmark of sauropod dinosaurs and is usually regarded as a key feeding adaptation². Here we describe a new dicraeosaurid sauropod, from the latest Jurassic period of Patagonia, that has a particularly short neck. With a neck that is about 40% shorter than in other known dicraeosaurids^{3,4}, this taxon demonstrates a trend opposite to that seen in most sauropods and indicates that the ecology of dicraeosaurids might have differed considerably from that of other sauropods. The new taxon indicates that there was a rapid radiation and dispersal of dicraeosaurids in the Late Jurassic of the Southern Hemisphere, after the separation of Gondwana from the northern continents by the late Middle Jurassic.

Dinosauria Owen, 1842

Saurischia Seeley, 1887

Sauropoda Marsh, 1878

Neosauropoda Bonaparte, 1986

Diplodocoidea (Marsh, 1884)

Dicraeosauridae Janensch, 1929

Brachytrachelopan mesai gen. et sp. nov.

Etymology. From *brachytrachelos* (Greek, short-necked), in reference to the short neck of the animal, and Pan (the Greek shepherd god), in reference to the fact that the specimen was found by Daniel Mesa while looking for stray sheep. The species name, *mesai*, honours Daniel Mesa and his family, who found the specimen.

Holotype. Museo Paleontológico Egidio Feruglio (Trelew) MPEF-PV 1716; an articulated partial skeleton, including 8 cervical, 12 dorsal, and 3 sacral vertebrae, the proximal parts of the posterior cervical and all dorsal ribs, the right ilium, distal part of the left femur, and proximal end of the left tibia.

Horizon and locality. The specimen comes from a fluvial sandstone within the Cañadón Cálcareo Formation⁵, Upper Jurassic (Tithonian). It was found on the summit of a hill about 25 km north-northeast of the village of Cerro Cándor. Exact locality data are available from the authors on request.

Diagnosis. *Brachytrachelopan mesai* differs from all other sauropods in its very short neck, with individual cervical vertebrae being as long as, or shorter in anteroposterior length than, high posteriorly. Further apomorphies of the taxon include a pronounced, pillar-like centropostzygapophyseal lamina in the cervical vertebrae, a pronounced anterior inclination of the mid-cervical neural spines, with the tip of the spine extending beyond the anterior end of the centrum, and anterior dorsal neural spines one to six with vertical bases and anteriorly flexed tips.

Description. *Brachytrachelopan mesai* is a small animal (by sauropod

standards) with an estimated total length of less than 10 m. Extensive fusion of the neural arches with their respective centra and fusion of sacral centra, sacral neural arches and sacral neural spines indicate that the animal does not represent a juvenile. The vertebral column from the fifth cervical to the third sacral vertebra is preserved in articulation with the ribs and the right ilium; the rest of the skeleton was lost to erosion, with the exception of fragments of the left femur and the tibia.

Eight cervical and 12 dorsal vertebrae are present; the total number of cervicals was probably 12, as in other dicraeosaurids^{6,7}. The cervical vertebrae are strongly opisthocoelous, whereas the dorsal and sacral vertebrae are amphiplatycelous. In comparison with other sauropod dinosaurs, the new taxon has a very short neck (Fig. 1). This is primarily a result of the shortening of the individual cervical vertebral centra, which are only as long as, or shorter anteroposteriorly (excluding the anterior convexity) than, their posterior height. As in *Dicraeosaurus*⁶, the cervical vertebrae have well-developed ventral keels and strongly constricted centra. Pleurocoels are absent from all preserved presacral vertebrae, and deep lateral depressions are present only in the last three cervical and two anteriormost dorsal vertebrae. As is typical in sauropods, the lateral laminae of the cervical neural arches are well developed. The centropostzygapophyseal lamina is developed as a massive column of bone. The neural spines are elongate rod-like, bifurcate and strongly inclined anteriorly in all preserved cervicals.

The cervical–dorsal transition is marked by an abrupt shift in the position of the parapophyses from the anteroventral end of the centrum in the last cervical to the anterodorsal end of the centrum in the first dorsal. The neural arches of the dorsal vertebrae are higher than their respective centra and have long, strongly dorsolaterally directed transverse processes. A posterior centroparapophyseal lamina is weakly developed in the first two dorsals and well developed from dorsal 8 to 12. A stout, ridge-like centropostzygapophyseal lamina is found in the infrapostzygapophyseal fossa in all dorsals. The dorsal neural spines are elongate and deeply bifurcated up to the sixth dorsal. The distal portion of each spine is markedly flexed anteriorly. The neural spine of the seventh dorsal is broad and petal-shaped, with only a slight dorsal indentation. As in other diplodocoids⁸, the neural spines of the posterior dorsals are mainly formed from combined prespinal, postspinal and spinopostzygapophyseal laminae and the lateral spinopostzygapophyseal laminae. A short spinodiapophyseal lamina is present and joins the spinopostzygapophyseal lamina in the basal third of the spine.

The two and a half sacral vertebrae that are preserved, and their neural arches, are fused and their neural spines form a continuous sheet of bone over the sacrum, which is supported laterally by the strongly developed lateral spinopostzygapophyseal laminae and the smaller but distinct spinodiapophyseal laminae.

Although the distal ends of the ribs are missing, the fully

¹Bayerische Staatssammlung für Paläontologie und Geologie, Richard-Wagner-Strasse 10, 80333 München, Germany. ²Museo Paleontológico Egidio Feruglio, Avenida Fontana 140, 9100 Trelew, Argentina.

articulated ribcage of *Brachytrachelopan* reveals details about ribcage morphology both in this animal and in sauropods in general. As in all diplodocoids^{1,9}, the cervical ribs are short and extend only slightly beyond the posterior end of their respective vertebrae. The transition from cervical to dorsal ribs is rather abrupt, with only the last cervical showing an intermediate rib morphology. All dorsal ribs, with the exception of the last element, are notably inclined posteriorly. There is a marked transition between the seventh and eighth dorsal vertebrae from thoracic ribs anteriorly to abdominal ribs posteriorly. Thoracic ribs are broad, with the distance between single elements being approximately equal to the width of the ribs, whereas abdominal ribs are approximately only half as broad and more widely spaced.

The appendicular elements present provide little anatomical information. The ilium, the distal end of the femur and the proximal end of the tibia seem to be generally similar to the corresponding elements in *Dicraeosaurus*¹⁰.

Although the Late Jurassic is usually regarded as the heyday of sauropod evolution, our knowledge of sauropods from this interval is based on only a few individual formations, almost all of which are in the Northern Hemisphere¹¹. The only notable Late Jurassic dinosaur site in the Southern Hemisphere known so far is the Tendaguru locality in Tanzania¹², which has yielded many sauropod remains. The sauropod assemblage from Tendaguru includes forms that are similar and closely related to the roughly contemporaneous taxa of the North American Morrison Formation^{3,13}, as well as several lineages unknown from the Late Jurassic of the Northern Hemisphere^{1,3}. The only lineage recorded from Tendaguru that is completely absent from the Laurasian fossil record is the dicraeosaurids. However, in the absence of other Late Jurassic Gondwanan sauropod faunas, it is impossible to evaluate the extent to which the Tendaguru assemblage is representative of Late Jurassic Gondwanan faunas. The Late Jurassic dinosaur fauna of Patagonia provides an opportunity to

test this. Phylogenetic analysis places *Brachytrachelopan* firmly within the Dicraeosauridae (Fig. 2; see Supplementary Information). Within this clade it represents the sister-group of the Late Jurassic African taxon *Dicraeosaurus* rather than *Amargasaurus* from the Lower Cretaceous of South America. Consequently, the new find indicates a rapid radiation and dispersal of dicraeosaurids after the separation of the continents of the Southern and Northern Hemispheres in the latest Middle Jurassic.

The most striking feature of *Brachytrachelopan* is its very short neck (Fig. 3). Generally, there is a tendency towards an increase in neck length in several lineages of sauropods, including basal forms, brachiosaurids, diplodocids and titanosaurs. This involves an increase in either the number of cervical vertebrae or the length of individual elements, or, more often, both. In extreme cases, neck length can reach almost 400% of the length of the dorsal vertebral column^{2,14}. The length of the neck of *Brachytrachelopan*, in contrast, is only 75% or less of that of the dorsal vertebral column, whereas this is 123% in *Dicraeosaurus*⁶ and 136% in *Amargasaurus*². Because other diplodocoids have relatively longer necks, and the longest neck within dicraeosaurids is found in the phylogenetically most basal taxon *Amargasaurus* (Fig. 2), there therefore seems to be a tendency towards shortening of the neck in dicraeosaurid phylogeny. The long neck of sauropods is usually regarded as one of the key feeding adaptations of this group², which greatly increased the feeding range of the animals in both the vertical and horizontal planes^{2,15–18}. The relative length, posture and flexibility of the neck seem to have been significant in niche partitioning in sauropods^{2,17–20}. Thus, the tendency towards neck shortening in dicraeosaurids indicates that these taxa were progressively adapting for low browsing and might have been specialized on specific food sources, as has been suggested for *Amargasaurus* and *Dicraeosaurus*^{2,19}. This is also supported by the structure of the cervical neural arches in *Brachytrachelopan*, which

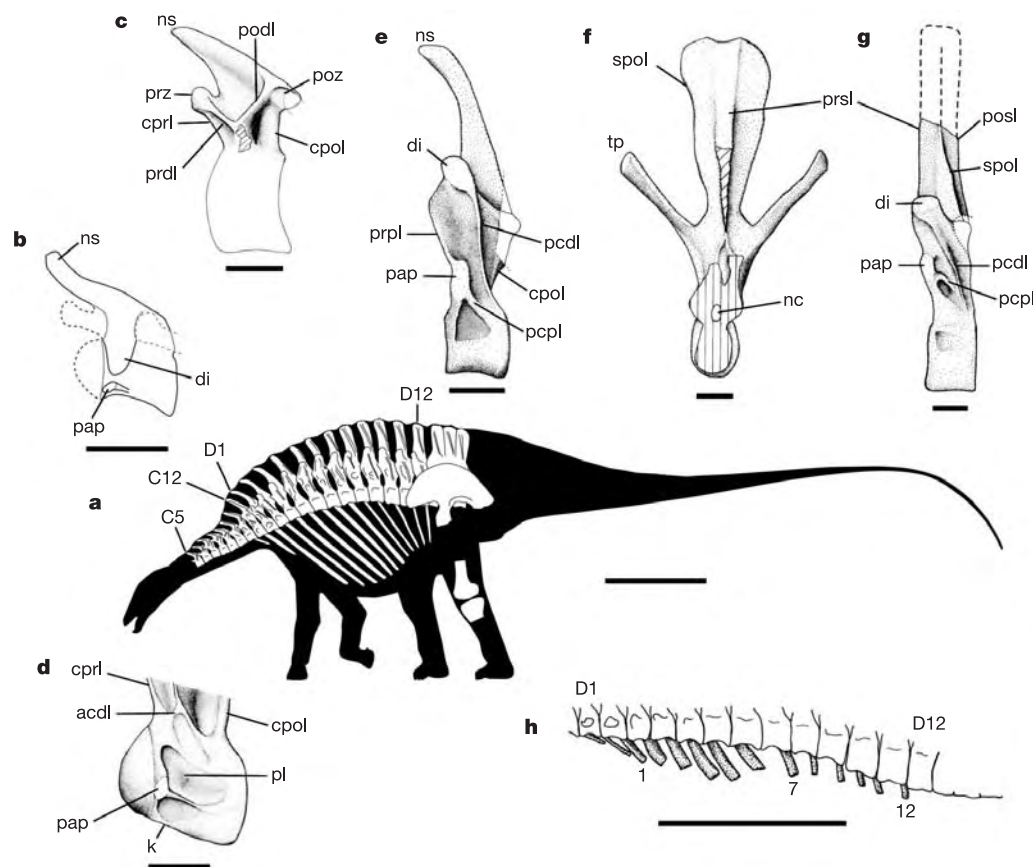


Figure 1 | Holotype of *Brachytrachelopan mesai*, MPEF PV 1716. a, Skeletal reconstruction showing preserved elements.

b, Reconstruction of 5th cervical vertebra in lateral view; c, neural arch of 8th cervical vertebra in lateral view; d, body of 10th cervical vertebra in lateral view; e, 2nd dorsal vertebra in lateral view; f, 7th dorsal vertebra in anterior view (body and neural arch sectioned); g, 11th dorsal vertebra in lateral view; h, vertebral column with articulated ribs of the right side as preserved. Scale bars, 1 m (a, h) and 10 cm (b–g). acdl, anterior centrodiapophyseal lamina; C, cervical vertebra; cpol, centropostzygapophyseal lamina; cpri, centroprezygapophyseal lamina; D, dorsal vertebra; di, diapophysis; k, ventral keel; nc, neural canal; ns, neural spine; pap, parapophysis; pcdl, posterior centrodiapophyseal lamina; pcpl, posterior centroparapophyseal lamina; pl, pleurocentral depression; podl, postzygodiapophyseal lamina; posl, postspinal lamina; poz, postzygapophysis; prdi, prezygodiapophyseal lamina; prpl, prezygoparapophyseal lamina; prsl, prespinal lamina; prz, prezygapophysis; spol, spinopostzygapophyseal lamina; tp, transverse process. Numbers 1, 7 and 12 in h refer to numbers of dorsal ribs.

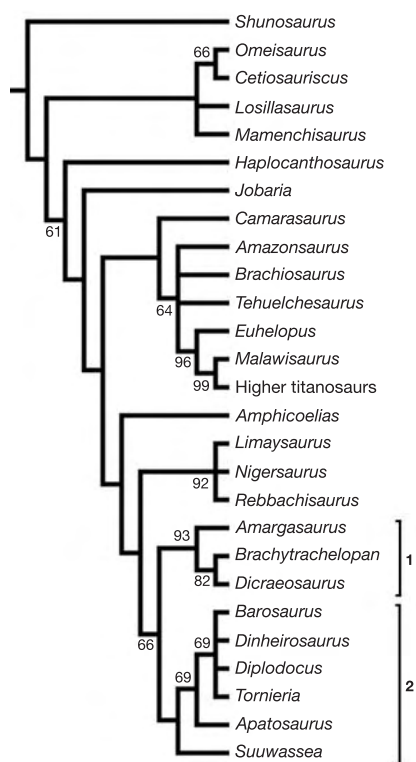


Figure 2 | Phylogenetic position of *Brachytrachelopan*, based on an analysis of 27 sauropod taxa and 154 characters. Numbers next to nodes indicate bootstrap support values for nodes that show more than 50% support. Bold numbers: 1, Dicraeosauridae; 2, Diplodocidae.



Figure 3 | Outline drawings of three diplodocoid sauropods for comparison of overall size and relative proportions. **a**, Diplodocid *Diplodocus carnegii* (Upper Jurassic Morrison Formation, USA); **b**, dicraeosaurid *Dicraeosaurus hansemanni* (Upper Jurassic Tendaguru Beds, Tanzania); **c**, dicraeosaurid *Brachytrachelopan mesai* (Upper Jurassic Cañadón Calcáreo Formation, Argentina). **a** and **b** are based on ref. 23.

would seriously have restricted dorsal flexion of the neck. Although this inference is somewhat hampered by a lack of forelimb material in *Brachytrachelopan*, it seems that this animal was specialized for feeding on plants growing at heights between about 1 and 2 m. Whereas most sauropods are generally regarded as obligatory bulk feeders because of their large body size and rather unspecialized feeding apparatus², this greater specialization might also have been a limiting factor for body size in dicraeosaurids, with all known taxa being at the lower end of the size range of sauropods¹. Thus, *Brachytrachelopan* might have had the same ecological role as large low-browsing iguanodontian ornithomimids in Late Jurassic ecosystems of the Northern Hemisphere, given its general similarity in size and body proportions to these forms, which seem to have been absent from Gondwanan Late Jurassic localities¹¹. It is interesting to note that the only other group of sauropods that are interpreted as specialist feeders is the sister-group of dicraeosaurids, the diplodocids². However, whereas dicraeosaurids became specialized feeders by shortening the neck and thus apparently limiting absolute body size, diplodocids solved this problem in a radically different way, by

greatly increasing neck length and thus potential feeding range². Thus, as already indicated by other studies of sauropod feeding mechanisms^{2,21,22}, sauropods may have shown greater adaptive plasticity than they are often given credit for.

Received 9 February; accepted 11 April 2005.

- Upchurch, P., Barrett, P. M. & Dodson, P. in *The Dinosauria* 2nd edn (eds Weishampel, D. B., Dodson, P. & Osmólska, H.) 259–322 (Univ. California Press, Berkeley, 2004).
- Upchurch, P. & Barrett, P. M. in *Evolution of Herbivory in Terrestrial Vertebrates. Perspectives from the Fossil Record* (ed. Sues, H.-D.) 79–122 (Cambridge Univ. Press, Cambridge, 2000).
- Janensch, W. Übersicht über die Wirbeltierfauna der Tendaguruschichten, nebst einer kurzen Charakterisierung der neu aufgeführten Arten von Sauropoden. *Archiv Biontol* 3, 81–110 (1914).
- Salgado, L. & Bonaparte, J. F. Un nuevo sauropodo Dicraeosauridae, *Amargasaurus cazaui* gen. et sp. nov., de la Formación La Amarga, Neocomiano de la Provincia del Neuquén Argentina. *Ameghiniana* 28, 333–346 (1991).
- Proserpio, C. A. Descripción geológica de la Hoja 44 e, Valle General Racedo, Pcia del Chubut. *Dirección Nacional de Minería y Geología Boletín* 201, 1–102 (1987).
- Janensch, W. Die Wirbelsäule der Gattung *Dicraeosaurus*. *Palaeontographica* 2 (suppl. 7), 35–133 (1929).
- Bonaparte, J. F. Evolución de las vértebras presacras en Sauropodomorpha. *Ameghiniana* 36, 115–187 (1999).
- Wilson, J. A. A nomenclature for vertebral laminae in sauropods and other saurischian dinosaurs. *J. Vertebr. Paleontol.* 19, 639–653 (1999).
- Wilson, J. A. Sauropod dinosaur phylogeny: critique and cladistic analysis. *Zool. J. Linn. Soc.* 136, 217–276 (2002).
- Janensch, W. Die Gliedmaßen und Gliedmaszengürtel der Sauropoden der Tendaguru-Schichten. *Palaeontographica* 3 (suppl. 7), 177–235 (1961).
- Weishampel, D. B., et al. in *The Dinosauria* 2nd edn (eds Weishampel, D. B., Dodson, P. & Osmólska, H.) 517–606 (Univ. of California Press, Berkeley, 2004).
- Maier, G. *African Dinosaurs Unearthed: The Tendaguru Expeditions* (Indiana Univ. Press, Bloomington, 2003).
- Remes, K. in *Geobiologie. 74. Jahrestagung der Paläontologischen Gesellschaft (Göttingen, 2–8th October 2004). Kurzfassungen der Vorträge und Poster* (eds Reitner, J., Reich, M. & Schmidt, G.) 195–196 (Universitätsdrucke Göttingen, Göttingen, 2004).
- Ouyang, H. & Ye, Y. *The First Mamenchisaurian Skeleton with Complete Skull: Mamenchisaurus youngi* (Sichuan Science and Technology Press, Chengdu, 2002).
- Bakker, R. T. Dinosaur feeding behaviour and the origin of flowering plants. *Nature* 274, 661–663 (1978).
- Martin, J. in *Fourth Symposium on Mesozoic Terrestrial Ecosystems, Short Papers* (eds Currie, P. J. & Koster, E. H.) 150–155 (Tyrrell Museum of Paleontology, Drumheller, Alberta, 1987).
- Stevens, K. A. & Parrish, J. M. Neck posture and feeding habits of two Jurassic sauropod dinosaurs. *Science* 284, 798–800 (1999).
- Christian, A. Neck posture and overall body design in sauropods. *Mitt. Mus. Naturkunde Berlin Geowissenschaftl. Reihe* 5, 271–281 (2002).
- Barrett, P. M. & Upchurch, P. in *Sixth Symposium on Mesozoic Terrestrial Ecosystems and Biota, Short Papers* (eds Sun, A. & Wang, Y.) 107–110 (China Ocean Press, Beijing, 1995).
- Foster, J. R. Relative abundance of the Sauropoda (Dinosauria, Saurischia) of the Morrison Formation and implications for Late Jurassic Paleogeography of North America. *Mesa Southw. Mus. Bull.* 8, 47–60 (2001).
- Calvo, J. O. Jaw mechanics in sauropod dinosaurs. *Gaia* 10, 183–193 (1994).
- Christiansen, P. Feeding mechanisms of the sauropod dinosaurs *Brachiosaurus*, *Camarasaurus*, *Diplodocus*, and *Dicraeosaurus*. *Historical Biol.* 14, 137–152 (2000).
- Paul, G. S. in *Dinofest International* (eds Wolberg, D. L., Stump, E. & Rosenberg, G. D.) 129–154 (Academy of Natural Sciences, Philadelphia, 1997).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank D. Mesa, who found the specimen, and his family for their hospitality during the fieldwork; all the preparators and volunteers who helped with the excavation and preparation of the material; H. Mayr for his help with nomenclatorial questions; and A. López-Arbarelo, P. Barrett and D. Unwin for critical comments. This project was supported by a German Academic Exchange Service (DAAD) postdoctoral fellowship, and grants to O.W.M.R. from the Jurassic Foundation, BBC *Horizon* and the German Research Foundation (DFG).

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to O.W.M.R. (o.rahut@lrz.uni-muenchen.de).

Oxytocin increases trust in humans

Michael Kosfeld^{1*}, Markus Heinrichs^{2*}, Paul J. Zak³, Urs Fischbacher¹ & Ernst Fehr^{1,4}

Trust pervades human societies^{1,2}. Trust is indispensable in friendship, love, families and organizations, and plays a key role in economic exchange and politics³. In the absence of trust among trading partners, market transactions break down. In the absence of trust in a country's institutions and leaders, political legitimacy breaks down. Much recent evidence indicates that trust contributes to economic, political and social success^{4,5}. Little is known, however, about the biological basis of trust among humans. Here we show that intranasal administration of oxytocin, a neuropeptide that plays a key role in social attachment and affiliation in non-human mammals^{6–8}, causes a substantial increase in trust among humans, thereby greatly increasing the benefits from social interactions. We also show that the effect of oxytocin on trust is not due to a general increase in the readiness to bear risks. On the contrary, oxytocin specifically affects an individual's willingness to accept social risks arising through interpersonal interactions. These results concur with animal research suggesting an essential role for oxytocin as a biological basis of prosocial approach behaviour.

In non-human mammals, the neuropeptide oxytocin has a central role in general behavioural regulation, particularly in positive social interactions. Aside from its well-known physiological functions in milk letdown and during labour, oxytocin receptors are distributed in various brain regions associated with behaviour^{9,10}, including pair bonding, maternal care, sexual behaviour, and the ability to form normal social attachments^{6–8,11–15}. Thus, oxytocin seems to permit animals to overcome their natural avoidance of proximity and thereby facilitates approach behaviour. Given that oxytocin is believed to promote social attachment and affiliation in non-human mammals, we hypothesized that oxytocin might also promote prosocial approach behaviours—such as trust—in humans. Recent research has shown that neuropeptides cross the blood-brain barrier after intranasal administration¹⁶, providing a useful method for studying the central nervous system effects of oxytocin in humans^{17,18}. We used a double-blind study design to compare trusting behaviour in a group of subjects that received a single dose of intranasal oxytocin with that of subjects in a control group that received placebo.

We analysed the effect of exogenously administered oxytocin on individuals' decisions in a trust game with real monetary stakes^{19–22}. In this trust game, two subjects interacting anonymously play either the role of an investor or a trustee (Fig. 1). First, the investor has the option of choosing a costly trusting action by giving money to the trustee. If the investor transfers money, the total amount available for distribution between the two players increases but, initially, the trustee reaps the whole increase. The trustee is then informed about the investor's transfer and can honour the investor's trust by sharing the monetary increase generated by the investor's transfer. Thus, if the investor gives money to the trustee and the latter shares the proceeds of the transfer, both players end up with a higher

monetary payoff. However, the trustee also has the option of violating the investor's trust. As sharing the proceeds is costly for the trustee, a selfish trustee will never honour the investor's trust because the investor and the trustee interact only once during the experiment.

The investor is therefore caught in a dilemma: if he trusts and the trustee shares, the investor increases his payoff, but he is also subject to the risk that the trustee will abuse this trust. In the latter case, the investor is worse off than if he had not trusted at all and, adding insult to injury, the trustee has an unfair payoff advantage relative to the investor. Substantial evidence exists to show that humans are averse to such risks^{22–24}. Moreover, the aversion of investors to abuse of trust seems to have an important role across different human cultures and social groups in the context of our game^{22,25}. The investors have to overcome their aversion against these risks in order to trust, allowing us to address the question of whether oxytocin modulates this trusting behaviour in humans.

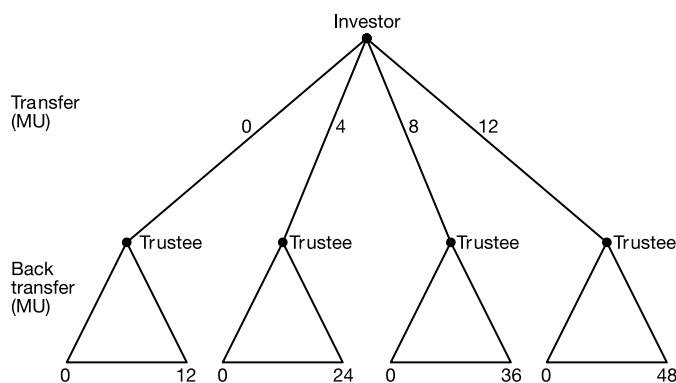


Figure 1 | The trust game. Both subjects receive an initial endowment of 12 monetary units (MU). The investor can send 0, 4, 8 or 12 MU to the trustee. The experimenter triples each MU the investor transfers. After the investor's decision is made, the trustee is informed about the investor's transfer. Then the trustee has the option of sending any amount between zero and his total amount available back to the investor. For example, if the investor has sent 12 MU, the trustee possesses 48 MU (12 MU own endowment + 36 MU tripled transfer) and can, therefore, choose any back transfer from 0 to 48 MUs. The experimenter does not triple the back transfer. The investor's final payoff corresponds to the initial endowment minus the transfer to the trustee, plus the back transfer from the trustee. The trustee's final payoff is given by his initial endowment plus the tripled transfer of the investor, minus the back transfer to the investor. At the end of the experiment, the earned MU are exchanged into real money according to a publicly announced exchange rate (see Methods). Each subject made four decisions in the same player role while paired with four different, randomly selected interaction partners.

¹University of Zurich, Institute for Empirical Research in Economics, Blumlisalpstrasse 10, CH-8006 Zurich, Switzerland. ²University of Zurich, Department of Clinical Psychology and Psychotherapy, Zurichbergstrasse 43, CH-8044 Zurich, Switzerland. ³Center for Neuroeconomics Studies, Claremont Graduate University, Claremont, California 91711-6165, USA. ⁴Collegium Helveticum, Schmelzbergstrasse 25, CH-8092 Zurich, Switzerland.

*These authors contributed equally to this work.

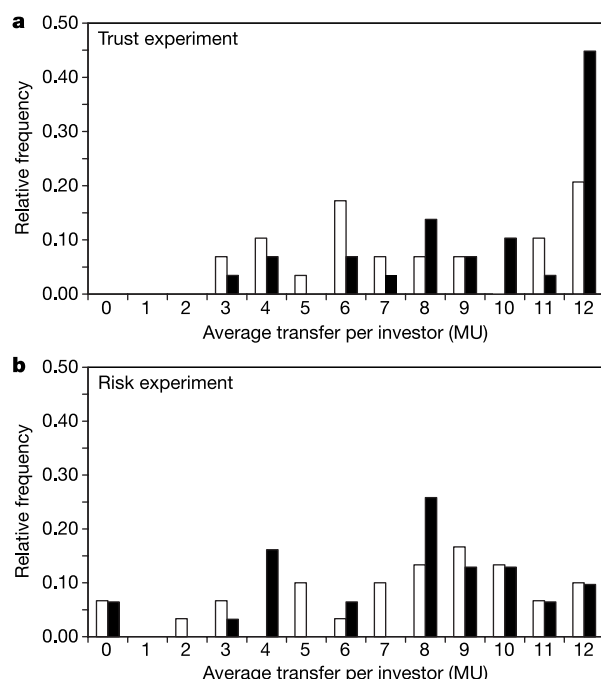


Figure 2 | Transfers in the trust and the risk experiment. Each observation represents the average transfer amount (in MU) over four transfer decisions per investor. **a**, Relative frequency of investors' average transfers in oxytocin (filled bars) and placebo (open bars) groups in the trust experiment ($n = 58$). Subjects given oxytocin show significantly higher transfer levels. **b**, Relative frequency of investors' average transfers in oxytocin (filled bars) and placebo (open bars) groups in the risk experiment ($n = 61$). Subjects in the oxytocin and the placebo group show statistically identical transfer levels.

Our hypothesis that oxytocin increases the trusting behaviour of investors implies that the investors in the oxytocin group ($n = 29$) will show higher money transfers than those in the placebo group ($n = 29$). In fact, our data show that oxytocin increases investors' trust considerably. Out of the 29 subjects, 13 (45%) in the oxytocin group showed the maximal trust level, whereas only 6 of the 29 subjects (21%) in the placebo group showed maximal trust (Fig. 2a). In contrast, only 21% of the subjects in the oxytocin group had a trust level below 8 monetary units (MU), but 45% of the subjects in the control group showed such low levels of trust. These differences in the distribution of trust result in higher average and median trust levels for subjects given oxytocin (Table 1). The investors' average transfer is 17% higher in the oxytocin group (Mann-Whitney U -test; $z = -1.897$, $P = 0.029$, one-sided), and the median transfer in the oxytocin group is 10 MU, compared to a median of only 8 MU for subjects in the placebo group.

In the trust game, the risk on the part of the investor's is due to the uncertainty of the trustee's behaviour—that is, a social interaction with a specific trustee constitutes the risk. This raises the question of whether oxytocin helps humans to overcome a general aversion

against risks or whether oxytocin specifically affects trusting behaviour in social interactions. In order to answer this question, we conducted a risk experiment in which the investor faced the same choices as in the trust game but in which a random mechanism, not the trustee's decision, determined the investor's risk. The random mechanism in the risk experiment replicated the trustees' decisions in the trust experiment. Therefore, the investors faced exactly the same risk as in the trust experiment (see Methods); however, their transfer decisions were not embedded in a social interaction because there were no trustees in the risk experiment.

In this risk experiment, the investors' behaviour does not differ between the oxytocin and the placebo groups (Table 1 and Fig. 2b). The median transfer is 8 MU and the average transfer is 7.5 MU in both groups (Mann-Whitney U -test; $z = 0.022$, $P = 0.983$; two-sided test, $n = 31$ in oxytocin group, $n = 30$ in placebo group). Moreover, there is no significant difference in a comparison of the placebo group in the trust experiment with the oxytocin group and the placebo group in the risk experiment (Kruskal-Wallis test; $\chi^2 = 0.533$, d.f. = 2, $P = 0.766$), with identical median transfers across groups (Table 1). However, if we add the oxytocin group in the trust experiment to these three samples, significant differences are observed (Kruskal-Wallis test; $\chi^2 = 8.610$, d.f. = 3, $P = 0.035$), indicating that only the investors in the oxytocin group of the trust experiment behave differently. Thus, oxytocin increases the investors' transfer levels in the trust experiment but not in the risk experiment. This finding is illustrated by a comparison of Figs 2a and b, which show that only 10% of the subjects with oxytocin choose the maximal transfer level in the risk experiment, whereas 45% choose the maximal level in the trust experiment. Therefore, the differences between the oxytocin group in the trust experiment and the oxytocin group in the risk experiment are highly significant (Mann-Whitney U -test; $z = -2.563$, $P = 0.010$, two-sided), suggesting that oxytocin specifically affects trust in interpersonal interactions.

The risk experiment constitutes a powerful control for the effects of oxytocin on trusting behaviour because everything is kept constant relative to the trust experiment, except that the investors' risk in the risk experiment is not generated through a social interaction. Specifically, all the indirect effects of oxytocin on the state of a subject, such as possible effects on mood or calmness, would be present in both the trust and the risk experiment. Therefore, these potential indirect effects of oxytocin cannot be responsible for the effect of oxytocin on trusting behaviour. Moreover, in order to provide an additional control for non-specific effects that might be associated with oxytocin administration, we explicitly measured mood and calmness before substance administration and 50 min after administration (but before subjects played the trust or the risk game). We used a questionnaire suitable for repeated measures within short periods of time, one that is widely used in neuropharmacological studies in humans²⁶ and correlates with physiological measures¹⁷. There were no statistical differences in the levels of mood and calmness before and after the administration of oxytocin in either the trust or the risk experiment. (Trust experiment: $z = -1.541$, $P = 0.123$ for calmness; $z = 1.452$, $P = 0.146$ for mood; $n = 29$. Risk experiment: $z = 0.620$, $P = 0.535$ for calmness; $z = -0.841$, $P = 0.400$ for mood; $n = 31$; two-sided Wilcoxon signed rank tests.) This provides further support for our conclusion

Table 1 | Median and average transfer behaviour of investors

	Trust experiment		Risk experiment	
	Oxytocin group	Placebo group	Oxytocin group	Placebo group
Mean average transfer (MU)	9.6	8.1	7.5	7.5
Median average transfer (MU)	10	8	8	8
Standard deviation of transfers (MU)	2.8	3.1	3.3	3.4
Number of observations	29	29	31	30

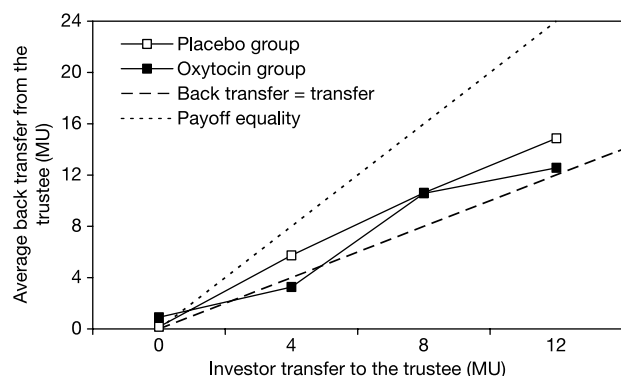


Figure 3 | Average back transfer of trustees to their investors. The graph shows the average back transfer by trustees for different levels of investor transfer in the oxytocin and placebo groups. The dotted line shows the level of the back transfer necessary to achieve payoff equality between the investor and the trustee. The dashed line shows a level of back transfer equal to the investor's transfer to the trustee. The trustees' back transfers are on average slightly higher than the amount sent by the investor. Trustees in both treatment groups make higher back transfers in response to higher original investor transfer levels. However, there is no statistically significant difference in back transfers between subjects in the oxytocin and the placebo groups.

that the effect of oxytocin on human trust is not caused by non-specific, psychotropic effects of oxytocin.

What mechanisms might be involved in generating the effect of oxytocin on trusting behaviour? One possibility is that oxytocin causes a general increase in prosocial inclinations. This implies that oxytocin should affect not only the prosocial behaviour of the investors but also that of the trustees. We would therefore predict that those trustees who are given oxytocin should make higher back transfers at any given level than the trustees who received placebo. However, trustees given oxytocin do not show more trustworthy behaviour (Fig. 3). At every positive transfer level (4, 8 or 12 MU), their back transfers are statistically indistinguishable from those of placebo trustees (Mann Whitney *U*-tests; $P > 0.243$, two-sided tests for each positive transfer level). Thus, oxytocin does not increase the general inclination to behave prosocially. Rather, oxytocin specifically affects the trusting behaviour of investors.

We hypothesize that the differing effect of oxytocin on the behaviour of investors and trustees is related to the fact that investors and trustees face rather different situations. Specifically, investors have to make the first step; they have to 'approach' the trustee by transferring money. In contrast, the trustees can condition their behaviour on the basis of the investors' actions. Thus, the psychology of trust is important for investors, whereas the psychology of strong reciprocity²⁷ is relevant for trustees. The fact that oxytocin affects subjects' approach or trust behaviour, but not their degree of reciprocity, is in agreement with animal studies. There is substantial evidence that oxytocin promotes prosocial approach behaviour by inhibiting defensive behaviours^{6,13}, but there is no evidence that oxytocin affects reciprocity in animals.

A second mechanism behind the effect of oxytocin on trust could be based on subjects' beliefs. Oxytocin might render subjects more optimistic about the likelihood of a good outcome. In order to address this question, we measured the investor's subjective expectation about the trustee's back transfer after every transfer decision. A Mann-Whitney *U*-test indicates that these expectations do not differ significantly between oxytocin and placebo groups at every feasible positive transfer level ($P > 0.357$, two-sided tests at transfer levels of 4, 8 or 12 MU). Thus, the investors given oxytocin show more trusting behaviour but do not hold significantly different beliefs about the trustworthiness of others. Moreover, oxytocin

does not affect investors' beliefs about the likelihood of a good outcome in the risk experiment ($P > 0.128$, two-sided Mann Whitney *U*-tests for transfer levels of 4, 8 or 12 MU).

Finally, there is the possibility that oxytocin helps subjects to overcome their betrayal aversion in social interactions. This explanation is consistent with the differing effects of oxytocin across the trust and the risk experiments, and is further supported by the fact that investors faced a considerable betrayal risk. An increase in the transfer level from 4 or 8 MU to 12 MU decreased the investor's average payoff slightly, whereas it increased the objective risk of very low back transfers by the trustee. However, betrayal aversion alone cannot explain why investors given oxytocin make higher transfers in the trust experiment compared with the risk experiment, because betrayal is impossible in the risk experiment. The higher transfers in the trust experiment can be reconciled with betrayal aversion if one acknowledges that investors' behaviour in the trust experiment is also likely to be driven by the motive to increase the available amount for distribution between the two players²⁸. As this motive cannot operate in the risk experiment, it can only increase transfers levels in the trust experiment. Our interpretation of oxytocin's effect on trust in terms of betrayal aversion may be seen in the light of animal studies indicating that increased availability of oxytocin in the central nervous system facilitates approach behaviour, by linking the overcoming of social avoidance with the activation of brain circuits implicated in reward (for example, the nucleus accumbens)^{12,15}.

The ubiquity of trusting behaviour is perhaps one of the distinguishing features of the human species. An element of trust characterizes almost all human social interactions. Here we have sought to examine the effect of oxytocin on trust in humans. Research in non-human mammals suggests that oxytocin has a key role in social attachment and affiliation. We find that intranasal administration of oxytocin causes a substantial increase in trusting behaviour. Subjects given oxytocin seem better able to overcome trust obstacles such as betrayal aversion. Of course, this finding could be misused to induce trusting behaviours that selfish actors subsequently exploit. However, our findings may also have positive clinical implications for patients with mental disorders that are associated with social dysfunctions (for example, social phobia or autism). In particular, social phobia ranks as the third most common mental health disorder and is characterized by marked social deficits, including persistent fear and avoidance of social interactions. Thus, our results might lead to fertile research on the role of oxytocin in several mental health disorders with major public health significance.

METHODS

Subjects. A total of 194 healthy male students (mean age $\pm$ s.d., 22.0 ± 3.4 yr) from different universities in Zurich participated in the study. The trust experiment had 128 participants, and 66 subjects participated in the risk experiment. Exclusion criteria for participation were significant medical or psychiatric illness, medication, smoking more than 15 cigarettes per day, and drug or alcohol abuse. Subjects were instructed to abstain from food and drink (other than water) for 2 h before the experiment, and from alcohol, smoking and caffeine for 24 h before the experiment. Participants were informed at the time of recruitment that the experiment would evaluate the effects of a hormone on decision making. In total, 16 individuals out of the original sample of 194 were excluded because of incorrect substance administration (7 in the trust experiment, 5 in the risk experiment) or their stated disbelief that the opponent in the trust game was actually a human being (4 participants). The study protocol was approved by the ethics committee of the University of Zurich. All subjects gave written, informed consent before participation.

Substance administration. Subjects received a single intranasal dose of 24 IU oxytocin (Syntocinon-Spray, Novartis; 3 puffs per nostril, each with 4 IU oxytocin) or placebo 50 min before the start of the trust or the risk experiment. Subjects were randomly assigned to the oxytocin or placebo group (double-blind, placebo-controlled study design). In order to avoid any subjective substance effects (for example, olfactory effects) other than those caused by oxytocin, the placebo contained all inactive ingredients except for the neuropeptide.

Behavioural experiment and questionnaires. After substance administration,

subjects completed questionnaires on a computer to measure demographic items and psychological characteristics. Owing to the crucial role of the social environment in triggering behavioural effects of oxytocin (as shown in animal research)^{13,29}, subjects were asked to wait in the rest area while the next part of the experiment was prepared. During this 5-min waiting period, subjects were seated at different tables. Subjects at the same table could talk to each other, but at the beginning of the experiment they were informed that they would not be interacting with those subjects who sat at the same table. When subjects re-entered the laboratory for both experiments, they received written instructions (available from the authors on request) explaining the payoff structure of the experiment and the private payment procedure at the end of the experiment. Subjects were randomly and anonymously assigned to the role of investor or trustee in the trust experiment, and did not know the identity of the persons with whom they were matched. After subjects had read the instructions in each experiment, we checked whether they understood the payoff structure by means of several hypothetical examples. All subjects (with one exception) answered the control questions correctly. One subject did not answer the control questions correctly and was excluded from the data set (this subject also did not apply the substance correctly). In addition, subjects received an oral summary of the instructions.

Each subject in the trust experiment made four decisions in the same player role while paired with different, randomly selected interaction partners. No pair of subjects interacted twice. Subjects in the role of the investor received no feedback about the trustee's decision between the different interactions. After every transfer decision, each investor was asked about his belief with regard to the expected back transfer from the trustee. Notably, trust levels were statistically constant across the four decisions. There is no time trend in investors' decisions in either the oxytocin or the placebo group. In the risk experiment, everything was identical to the trust experiment, except that all subjects played the role of an investor who could transfer 0, 4, 8, or 12 MU into a project rather than to a trustee. In particular, an investor's payoff risk (that is, the distribution of payoffs) in the risk experiment was identical to that in the trust experiment at any feasible transfer level.

To measure alterations in the psychological state of subjects throughout the course of the experiment, we assessed their mood and calmness at the beginning of the experiment (before substance administration) and immediately before the trust experiment or the risk experiment, by means of a suitable questionnaire²⁶. All decisions in the experiments and the answers to the questionnaires were entered on a computer using z-Tree software³⁰. Subjects received a flat fee of 80 Swiss francs for participation in the experiment; each MU earned in the trust and the risk experiment was worth 0.40 Swiss francs.

Received 20 April; accepted 5 May 2005.

- Luhmann, N. *Trust and Power* (Wiley, New York, 1979).
- Coleman, J. *Foundations of Social Theory* 91–118 (Belknap Press of Harvard Univ. Press, Cambridge, 1990).
- Arrow, K. Gifts and exchanges. *Phil. Publ. Affairs* 1, 343–362 (1972).
- Knack, S. & Keefer, P. Does social capital have an economic payoff? A cross-country investigation. *Q. J. Econ.* 112, 1251–1288 (1997).
- Zak, P. J. & Knack, S. Trust and growth. *Econ. J.* 111, 295–321 (2001).
- Carter, C. S. Neuroendocrine perspectives on social attachment and love. *Psychoneuroendocrinology* 23, 779–818 (1998).
- Uvnas-Moberg, K. Oxytocin may mediate the benefits of positive social interaction and emotions. *Psychoneuroendocrinology* 23, 819–835 (1998).
- Insel, T. R. & Young, L. J. The neurobiology of attachment. *Nature Rev. Neurosci.* 2, 129–136 (2001).
- Landgraf, R. & Neumann, I. D. Vasopressin and oxytocin release within the brain: a dynamic concept of multiple and variable modes of neuropeptide communication. *Front. Neuroendocrinol.* 25, 150–176 (2004).

- Huber, D., Pierre, V. & Ron, S. Vasopressin and oxytocin excite distinct neuronal populations in the central amygdala. *Science* 308, 245–248 (2005).
- Carter, C. S., Altemus, M. & Chrousos, G. P. Neuroendocrine and emotional changes in the post-partum period. *Prog. Brain Res.* 133, 241–249 (2001).
- Young, L. J., Lim, M. M., Gingrich, B. & Insel, T. R. Cellular mechanisms of social attachment. *Horm. Behav.* 40, 133–138 (2001).
- Pedersen, C. A. Oxytocin control of maternal behavior. Regulation by sex steroids and offspring stimuli. *Ann. NY Acad. Sci.* 807, 126–145 (1997).
- Heinrichs, M., Neumann, I. & Ehler, U. Lactation and stress: protective effects of breast-feeding in humans. *Stress* 5, 195–203 (2002).
- Insel, T. R. & Shapiro, L. E. Oxytocin receptor distribution reflects social organization in monogamous and polygamous voles. *Proc. Natl Acad. Sci. USA* 89, 5981–5985 (1992).
- Born, J. et al. Sniffing neuropeptides: a transnasal approach to the human brain. *Nature Neurosci.* 5, 514–516 (2002).
- Heinrichs, M., Baumgartner, T., Kirschbaum, C. & Ehler, U. Social support and oxytocin interact to suppress cortisol and subjective responses to psychosocial stress. *Biol. Psychiatry* 54, 1389–1398 (2003).
- Heinrichs, M., Meinlschmidt, G., Wippich, W., Ehler, U. & Hellhammer, D. H. Selective amnesic effects of oxytocin on human memory. *Physiol. Behav.* 83, 31–38 (2004).
- Camerer, C. & Weigelt, K. Experimental tests of a sequential equilibrium reputation model. *Econometrica* 56, 1–36 (1988).
- Fehr, E., Kirchsteiger, G. & Riedl, A. Does fairness prevent market clearing? An experimental investigation. *Q. J. Econ.* 108, 437–459 (1993).
- Berg, J., Dickhaut, J. & McCabe, K. Trust, reciprocity and social history. *Games Econ. Behav.* 10, 122–142 (1995).
- Bohnet, I. & Zeckhauser, R. Trust, risk and betrayal. *J. Econ. Behav. Organ.* 55, 467–484 (2004).
- Holt, C. & Laury, S. Risk aversion and incentive effects. *Am. Econ. Rev.* 92, 1644–1655 (2002).
- Fehr, E. & Schmidt, K. M. A theory of fairness, competition, and cooperation. *Q. J. Econ.* 114, 817–868 (1999).
- Hong, K. & Bohnet, I. *Status and Distrust: the Relevance of Inequality and Betrayal Aversion* (Working Paper RW04-041, Kennedy School, Harvard Univ., Cambridge, 2004).
- Steyer, R., Schwenkmezger, P., Notz, P. & Eid, M. *Der Mehrdimensionale Befindlichkeitsfragebogen (MDBF)* [Multidimensional mood questionnaire] (Hogrefe, Göttingen, 1997).
- Gintis, H., Bowles, S., Boyd, R. & Fehr, E. Explaining altruistic behavior in humans. *Evol. Hum. Behav.* 24, 153–172 (2003).
- Engelmann, D. & Strobel, M. Inequality aversion, efficiency, and maximin preferences in simple distribution experiments. *Am. Econ. Rev.* 94, 857–869 (2004).
- Kendrick, K. M. et al. Neural control of maternal behaviour and olfactory recognition of offspring. *Brain Res. Bull.* 44, 383–395 (1997).
- Fischbacher, U. z-Tree. *Zurich Toolbox for Readymade Economic Experiments* (Working Paper No. 21, Institute for Empirical Research in Economics, Univ., Zurich, 1999).

Acknowledgements This paper is part of the Research Priority Program 'Foundations of Human Social Behaviour—Altruism versus Egoism' at the University of Zurich. Financial support from the MacArthur Foundation (Network on Economic Environments and the Evolution of Individual Preferences and Social Norms) and the Cogito Foundation is gratefully acknowledged. M.H. also acknowledges support from the Swiss National Science Foundation. We thank F. Heusi, A. Enzler, S. Gilomen, E. Götz, M. Koenigsberg, K. Korsunsky, S. Krammer, S. Lauber, P. Obrist-Rybar, M. Schellhammer, B. Studer, R. Stünzi and M. Trottman for research assistance.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to E.F. (efehr@iew.unizh.ch) and M.H. (m.heinrichs@psychologie.unizh.ch).

An inhibitor of Bcl-2 family proteins induces regression of solid tumours

Tilman Oltersdorf^{1*}, Steven W. Elmore^{2*}, Alexander R. Shoemaker^{2*}, Robert C. Armstrong¹, David J. Augeri², Barbara A. Belli¹, Milan Bruncko², Thomas L. Deckwerth¹, Jurgen Dinges², Philip J. Hajduk², Mary K. Joseph², Shinichi Kitada³, Stanley J. Korsmeyer^{4,5}, Aaron R. Kunzer², Anthony Letai⁵, Chi Li⁶, Michael J. Mitten², David G. Nettesheim², ShiChung Ng², Paul M. Nimmer², Jacqueline M. O'Connor², Anatol Oleksijew², Andrew M. Petros², John C. Reed³, Wang Shen², Stephen K. Tahir², Craig B. Thompson⁶, Kevin J. Tomaselli¹, Baole Wang², Michael D. Wendt², Haichao Zhang², Stephen W. Fesik² & Saul H. Rosenberg²

Proteins in the Bcl-2 family are central regulators of programmed cell death¹, and members that inhibit apoptosis, such as Bcl-X_L and Bcl-2, are overexpressed in many cancers and contribute to tumour initiation, progression and resistance to therapy². Bcl-X_L expression correlates with chemo-resistance of tumour cell lines³, and reductions in Bcl-2 increase sensitivity to anticancer drugs⁴ and enhance *in vivo* survival⁵. The development of inhibitors of these proteins as potential anti-cancer therapeutics has been previously explored^{6–15}, but obtaining potent small-molecule inhibitors has proved difficult owing to the necessity of targeting a protein–protein interaction. Here, using nuclear magnetic resonance (NMR)-based screening, parallel synthesis and structure-based design, we have discovered ABT-737, a small-molecule inhibitor of the anti-apoptotic proteins Bcl-2, Bcl-X_L and Bcl-w, with an affinity two to three orders of magnitude more potent than previously reported compounds^{7–15}. Mechanistic studies reveal that ABT-737 does not directly initiate the apoptotic process, but enhances the effects of death signals, displaying synergistic cytotoxicity with chemotherapeutics and radiation. ABT-737 exhibits single-agent-mechanism-based killing of cells from lymphoma and small-cell lung carcinoma lines, as well as primary patient-derived cells, and in animal models, ABT-737 improves survival, causes regression of established tumours, and produces cures in a high percentage of the mice.

Anti-apoptotic family members (for example, Bcl-2 and Bcl-X_L) and a subgroup of pro-apoptotic proteins (for example, Bax and Bak) are α -helical proteins with extensive sequence and structural similarity¹⁶. A separate pro-apoptotic group has similarity restricted to a single α -helix called the BH3 region (for example, Bad and Bid). These BH3-only proteins initiate apoptosis either by activating pro-apoptotic Bcl-2 proteins or by inhibiting the anti-apoptotic family members^{17,18}. Their activity is mediated through the association of the BH3 α -helix of one protein with a large hydrophobic pocket on binding partners¹⁹.

A high-throughput NMR-based method for lead compound discovery called 'SAR by NMR'²⁰ was used to screen a chemical library to identify small molecules that bind to the hydrophobic BH3-binding groove of Bcl-X_L. 4'-Fluoro-biphenyl-4-carboxylic acid (dissociation constant $K_d = 0.30 \pm 0.03$ mM; mean $\pm$ s.d., $n = 3$) and 5,6,7,8-tetrahydro-naphthalen-1-ol ($K_d = 4.3 \pm 1.6$ mM) bind to distinct

but proximal subsites within this cleft. The carboxyl group of the 4-biphenylcarboxylic acid binds near Arg 139 of Bcl-X_L, while the 4'-fluorophenyl group occupies a hydrophobic pocket (site 1) formed by Tyr 101, Leu 108, Val 126 and Phe 146 (Fig. 1a). Interestingly, these are the same sites that are occupied by Asp 83 and Leu 78 of a peptide derived from the BH3 region of Bak, two of the three residues most critical for affinity to Bcl-X_L (ref. 19). Similarly, the naphthalene derivative occupies the same hydrophobic region (site 2) as the third critical residue of the Bak peptide, Ile 85.

The SAR by NMR technology is based on the linkage of proximal fragments to achieve high-affinity binding. Substitution of an acyl-sulphonamide for the biphenyl carboxyl group maintained the correct positioning of the acidic proton while providing an optimal trajectory to site 2 that avoids steric interference by Phe 97. Site-directed parallel synthesis led to the discovery that a 3-nitro-4-(2-phenylthioethyl)aminophenyl group spans the binding sites and efficiently occupies site 2 through hydrophobic collapse and subsequent π - π stacking (Fig. 1b, compound 1).

Compound 1 binds with high affinity to Bcl-X_L (inhibitory constant $K_i = 36 \pm 1.6$ nM; mean $\pm$ s.e.m., $n = 3$), but affinity was attenuated by a factor of >280 in the presence of 1% human serum owing to tight binding to domain III of human serum albumin (HSA). To reduce binding to HSA, a structure-based approach was used²¹. The NMR structure of the thioethylamino-2,4-dimethylphenyl analogue of compound 1 complexed with domain III of HSA (Fig. 1c) was compared to the structure of compound 1 bound to Bcl-X_L. The structures revealed portions of the molecule that were solvent-exposed in the Bcl-X_L complex, but surrounded by lipophilic residues in the complex with HSA. These positions (arrows in Fig. 1c) were modified with polar substituents to reduce binding to HSA without affecting affinity for Bcl-X_L. Specifically, a basic 2-dimethylaminoethyl group was appended to the 1-position of the thioethylamino linkage group and the fluorophenyl occupying site 1 was replaced with a substituted piperazine. In addition, to improve the binding to Bcl-2, a lipophilic group was added to the piperazine to access a deep, well-defined pocket identified in three-dimensional structures of Bcl-2/inhibitor complexes (see Supplementary Fig. 2).

The resultant compound, ABT-737 (Fig. 1d), binds with high affinity ($K_i \leq 1$ nM) to Bcl-X_L, Bcl-2 and Bcl-w, but not to the less-homologous proteins Bcl-B, Mcl-1 and A1 ($K_i = 0.46 \pm 0.11$ μ M,

¹Idun Pharmaceuticals, 9380 Judicial Drive, San Diego, California 92121, USA. ²Global Pharmaceutical Research & Development, Abbott Laboratories, Abbott Park, Illinois 60064, USA. ³The Burnham Institute, 10901 North Torrey Pines Road, La Jolla, California 92037, USA. ⁴Howard Hughes Medical Institute, Harvard Medical School, Boston, Massachusetts 02115, USA. ⁵Harvard Medical School, Dana-Farber Cancer Institute, Boston, Massachusetts 02115, USA. ⁶Abramson Family Cancer Research Institute, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA.

*These authors contributed equally to this work.

>1 μM and >1 μM , respectively). Nanomolar activity was retained in the presence of 10% human serum (median inhibitory concentration $\text{IC}_{50} = 35 \pm 1 \text{ nM}$ and $103 \pm 2 \text{ nM}$ for Bcl- X_L and Bcl-2, respectively). The enantiomer (the stereoisomer bearing the opposite configuration of the dimethylaminoethyl group) was less potent ($\text{IC}_{50} = 1,400 \pm 150$ and $650 \pm 70 \text{ nM}$ for Bcl- X_L and Bcl-2, respectively, 10% human serum) and was employed as a loss-of-function control in the biological assays described below.

'Activating' BH3 proteins (for example, Bid, Bim) initiate apoptosis by inducing Bax and Bak oligomerization and subsequent cytochrome *c* release, while 'sensitizing' BH3 proteins (for example, Bad and Bik) exert their pro-death function by binding anti-apoptotic Bcl-2 family members, thereby preventing them from sequestering activating BH3 proteins²². Thus Bid protein induced cytochrome *c* release from purified mitochondria, which was inhibited by Bcl-2, while a Bad-derived BH3 peptide did not cause cytochrome *c* release by itself (Fig. 2a). However, the Bad BH3 peptide blocked the inhibition of Bid-mediated cytochrome *c* release by Bcl-2. ABT-737 showed the same effect as a Bad-derived BH3 peptide in purified mitochondria. Like the Bad BH3 peptide (but in contrast to a stabilized peptide derived from Bid¹⁴), ABT-737 antagonized Bcl-2 protection at concentrations $\geq 10 \text{ nM}$, but did not induce cytochrome *c* release at concentrations up to 1 μM . The enantiomer was significantly less active. ABT-737 also reversed the protection afforded by Bcl- X_L , activity that was completely dependent on Bax or Bak (see Supplementary Fig. 3)²³. Moreover, the presence of either of these pro-apoptotic proteins proved sufficient to mediate the effects of ABT-737. Thus, like Bad, ABT-737 binds to and inhibits anti-apoptotic Bcl-2 family proteins, but does not directly activate the pro-apoptotic proteins Bax and Bak.

A mammalian two-hybrid system²⁴ demonstrated that ABT-737 disrupted an intracellular Bcl-2 family protein–protein interaction.

ABT-737 inhibited the interaction of GAL4–Bcl- X_L and VP16–Bcl- X_S by $44 \pm 1\%$ and $55 \pm 2\%$ at concentrations of 0.1 μM and 1.0 μM , respectively, while the enantiomer exhibited little activity ($5 \pm 1\%$ inhibition at 3 μM). The mechanism was further validated by confocal time-lapsed microscopy experiments, in which ABT-737 (but not the enantiomer) displaced a GFP-tagged BH3-only protein (Bcl- G_S) from Bcl- X_L localized at mitochondrial surfaces in intact tumour cells (Fig. 2b–g).

ABT-737 and related compounds display synergism with chemotherapeutics and radiation, such that the median effective concentration (EC_{50}) value for cytotoxicity of the chemotherapeutic/radiation is reduced in the presence of ABT-737 (etoposide, 2–13-fold; doxorubicin, 2–4-fold; cisplatin, 2–3-fold; paclitaxel 2–20-fold; radiation, 2–4-fold) in a variety of tumour cell lines (not shown).

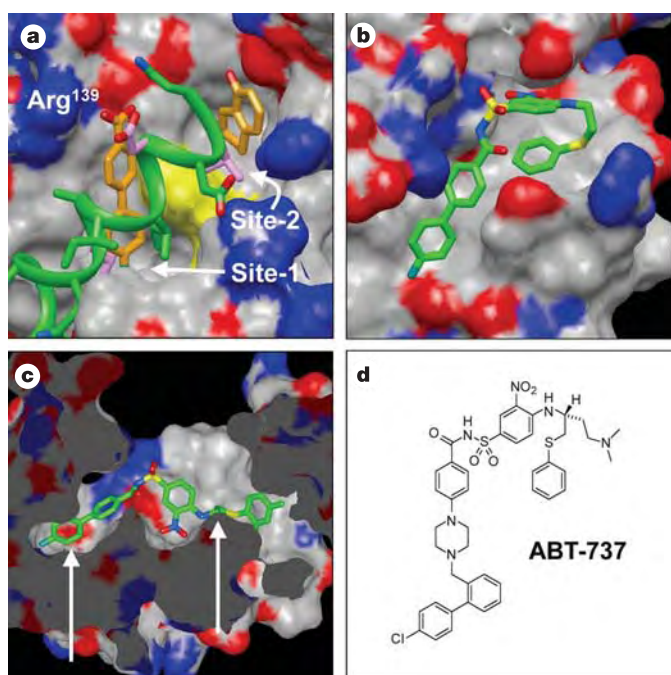
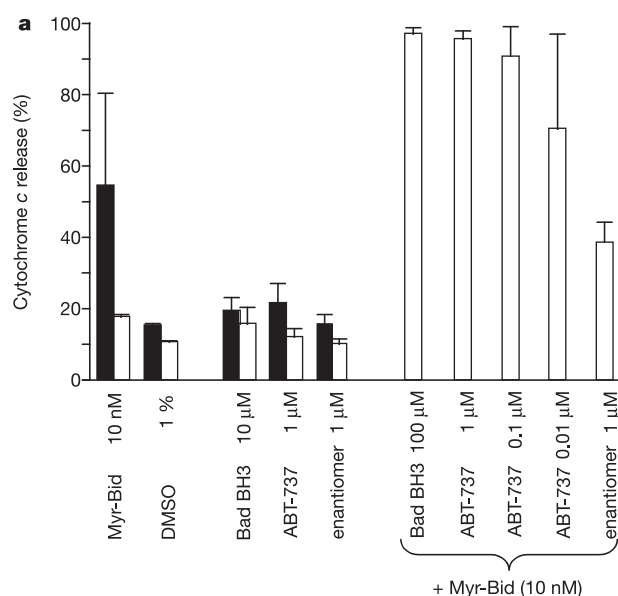


Figure 1 | Generation of ABT-737. **a**, Connolly surface of the complex of Bcl- X_L with a Bak-derived peptide (GQVGRQLAIIGDDINR, green) overlaid on the ternary complex of Bcl- X_L and NMR-based screening leads (orange). Phe 97 is shown in yellow. Residues of the Bak peptide (Asp 83, Leu 78, Ile 85) critical for binding are shown in magenta. **b**, Connolly surface of the NMR structure of Bcl- X_L complexed with compound 1. **c**, Cutaway Connolly surface for the NMR structure of domain III of HSA bound to the thioethylamino-2,4-dimethylphenyl analogue of compound 1. Arrows indicate proposed sites of modification. **d**, Chemical structure of ABT-737.

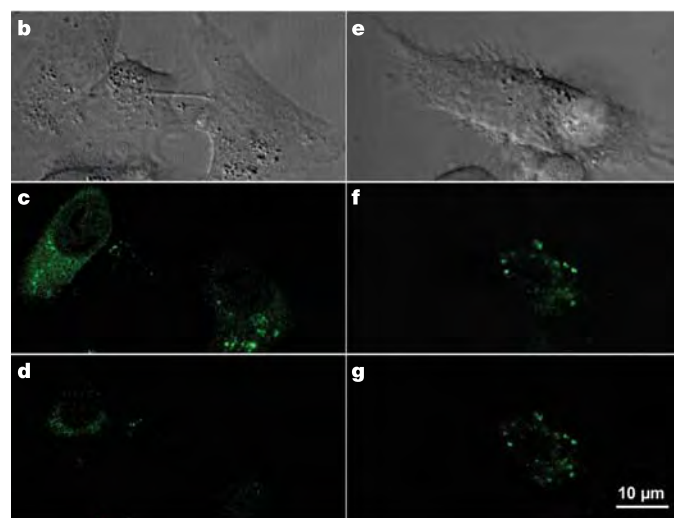


Figure 2 | ABT-737 antagonizes anti-apoptotic Bcl-2 family proteins. **a**, ABT-737 and Bad BH3 peptide do not induce cytochrome *c* release from wild-type (filled bar) or Bcl-2 transfected (open bar) mitochondria, but antagonize the ability of Bcl-2 to inhibit Bid-mediated cytochrome *c* release. Data are mean $\pm$ s.d. **b–g**, Confocal microscopic images of HeLa cells transfected with Bcl- X_L and GFP-Bcl- G_S visualized with GFP fluorescence. **b**, **e**, Phase-contrast images of untreated cells. **c**, **f**, Bcl- G_S fused to GFP displays a punctate pattern coincident with mitochondria. **d**, **g**, ABT-737 (100 nM, **d**) but not the enantiomer (100 nM, **g**) displaces GFP-labelled Bcl- G_S from mitochondria.

This synergism suggests that Bcl-2 family inhibitors may be useful as a component of several chemotherapy regimens. For example, ABT-737 enhanced the cytotoxicity of paclitaxel against A549 NSCLC (non small-cell lung carcinoma) cells by a factor of 4 (Fig. 3a), although ABT-737 alone did not affect cell survival at concentrations up to 1 μ M. Because single-agent anti-tumour activity is a more readily achieved clinical endpoint than chemosensitization, we determined whether any tumour cells would be sensitive to our Bcl-2 family inhibitors as single agents. Thus we evaluated the effects of ABT-737 against a diverse panel of human tumour cell lines. In contrast to the weak activity observed with many solid tumour lines such as A549 cells, ABT-737 displayed potent single-agent activity against the subset of cell lines representing lymphoid malignancies and SCLC.

Defective apoptosis is frequently associated with malignancies originating from B-lymphocytes. Follicular lymphoma, the most common B-cell neoplasm, is characterized by overexpression of *bcl-2* through a t(14;18)(q32;q21) chromosomal translocation²⁵. ABT-737 by itself was cytotoxic against the t(14;18) chromosomal

translocation-containing lymphoma cell lines RS11380, DoHH2 and SuDHL-4 (median effective concentration $EC_{50} = 0.15 \pm 0.03 \mu$ M, $0.13 \pm 0.01 \mu$ M, and $0.85 \pm 0.14 \mu$ M, respectively; mean $\pm$ s.e.m., $n = 3$). The enantiomer was weaker by a factor of >29 , suggesting that the activity was the direct result of binding to Bcl-X_L or Bcl-2. Moreover, ABT-737 significantly improved survival in a tumour model of disseminated disease using DoHH-2 cells (see Supplementary Fig. 4), confirming the activity observed in tissue culture.

To demonstrate that the cytotoxicity was not restricted to established tumour cell lines, we evaluated ABT-737 against primary patient-derived follicular lymphoma cells. ABT-737 (10–100 nM), but not the enantiomer, induced apoptosis in these samples (Fig. 3b). Similarly, ABT-737 induced concentration-dependent apoptosis in 13 of 15 patient-derived chronic lymphocytic leukaemia (CLL) B-cell specimens (Fig. 3c), while the enantiomer control had little effect at concentrations up to 100 nM.

Activity against SCLC-derived cells suggests that ABT-737 could also be useful for the treatment of solid tumours. The majority of SCLC cell lines studied were sensitive to ABT-737 ($EC_{50} < 1 \mu$ M),

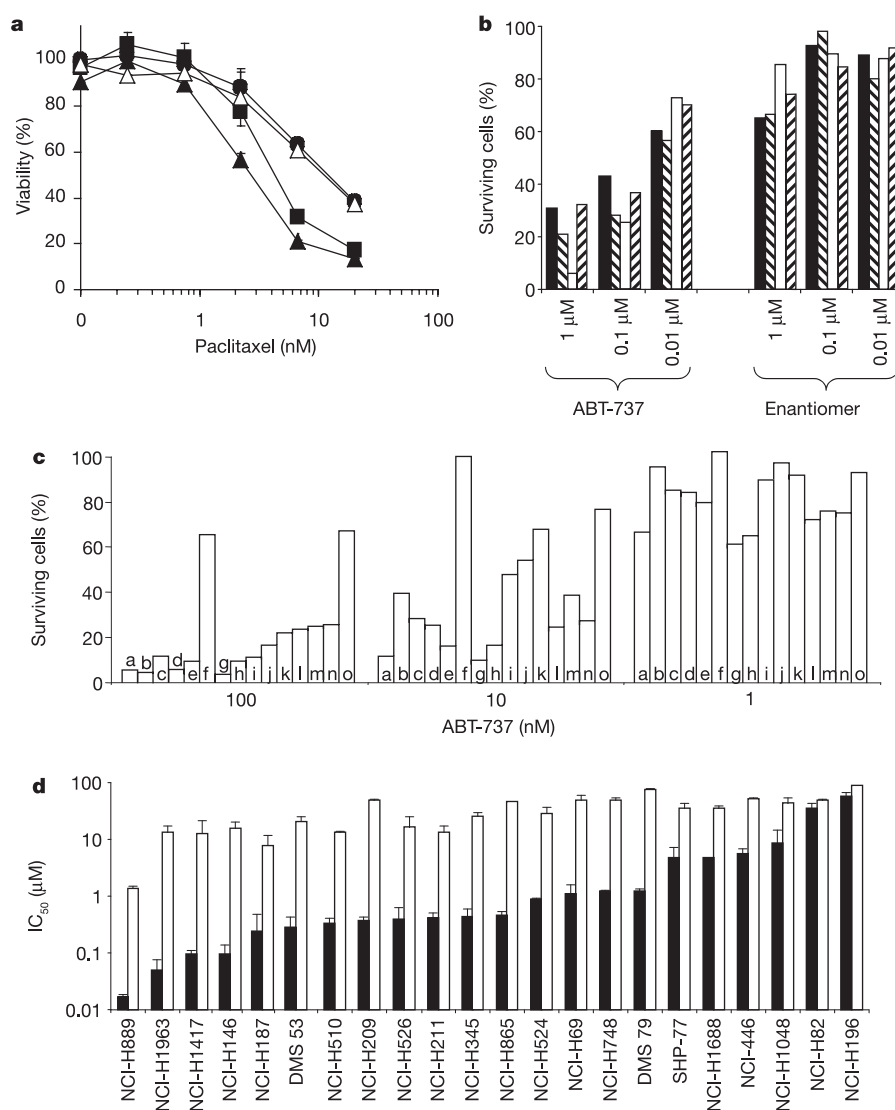


Figure 3 | Cell-based activity of ABT-737. **a**, ABT-737 enhances the activity of paclitaxel against A549 NSCLC cells. Filled circle, vehicle; filled square and filled triangle, 0.37 μ M and 1.1 μ M ABT-737, respectively; open triangle, 1.1 μ M enantiomer. Data are mean $\pm$ s.d. **b**, **c**, ABT-737 induces apoptosis of primary follicular lymphoma cells (**b**) and primary CLL cells (**c**). Values are normalized to survival of untreated cells (bar pattern (**b**) and letters within

bars (**c**) designate individual patient samples). CLL cell survival following enantiomer treatment is $>76\%$ and $>87\%$ at 100 nM and 10 nM, respectively. **d**, ABT-737 (filled bar), but not the enantiomer (open bar), is active against SCLC cell lines cultured in the presence of 10% human serum. Data are mean $\pm$ s.d.

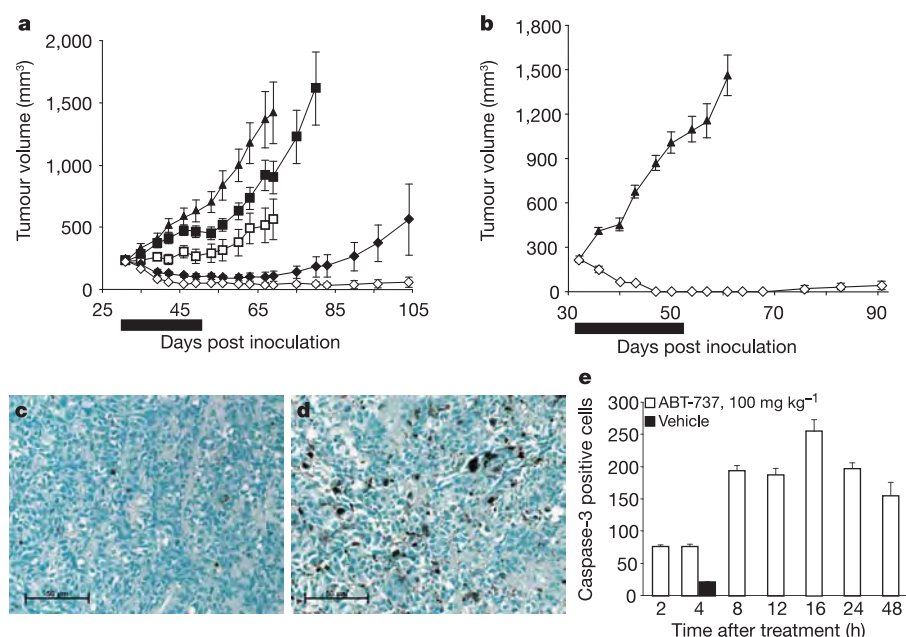


Figure 4 | *In vivo* anti-tumour activity of ABT-737. **a, b**, ABT-737 is efficacious in the H146 (**a**) and H1963 (**b**) SCLC xenograft models. Filled triangle, vehicle; filled square, ABT-737 administered i.p., 25 mg kg⁻¹ day⁻¹; open square, 50 mg kg⁻¹ day⁻¹; filled diamond, 75 mg kg⁻¹ day⁻¹; open diamond, 100 mg kg⁻¹ day⁻¹ (complete regression of 77% H1963 tumours, and 77% and 20% of H146 tumours treated at 100 and 75 mg kg⁻¹ day⁻¹,

respectively. Regression was also observed in H889, H1417, H187, H128 and H345 xenograft models, data not shown). Black bar indicates dosing period, data are mean $\pm$ s.e.m. **c–e**, ABT-737 induces apoptosis *in vivo*. Caspase-3 staining of vehicle (<5% of total tumour cells, 4 h after administration) (**c**) or ABT-737 (100 mg kg⁻¹, 16 h after administration) (**d**) treated H146 tumours. **e**, Time course for induction of apoptosis. Data are mean $\pm$ s.e.m.

and the >30-fold reduced potency of the enantiomer supports Bcl-2/X_L-mediated activity (Fig. 3d). Furthermore, in the *in vivo* setting, ABT-737 caused complete regression of established SCLC tumour xenografts (Fig. 4a and b), and tumours did not grow back in a high percentage of the mice for the duration of the study (58 days or 107 days after the cessation of therapy for H146 and H1963 mouse models, respectively).

To demonstrate that tumour regression was the result of apoptotic cell death, established H146 tumours were treated with a single dose of ABT-737 and analysed with an antibody specific for activated caspase-3 (Fig. 4c–e). A significant increase in caspase-3-positive cells was noted as early as 2 h after treatment, with a 12-fold increase achieved within 16 h. Examination of liver, heart and intestine revealed no increase in caspase-3 activation in these normal tissues (data not shown). Furthermore, ABT-737 was well tolerated by all observable measures and did not produce significant weight loss (<5%) in mice treated for 21 days, although clinical pathology revealed a reduction in platelets and lymphocytes. Taken together, our results suggest that Bcl-2 family inhibitors may be useful for the treatment of lymphoma and SCLC as monotherapy and a wide variety of cancers when given in combination with chemotherapy or radiation.

METHODS

NMR-based screening and structural studies. Bcl-X_L, Bcl-2 and domain III of HSA were prepared as previously described^{19,26,27}. NMR data were collected at 303 K on Bruker DRX500, DRX600 and DRX800 spectrometers. NMR-based screening was conducted as previously described²⁰. Intermolecular and intra-protein distance restraints were obtained from three-dimensional ¹⁵N- and ¹³C-nuclear Overhauser effect spectroscopy (NOESY) data, and structure calculations were performed using a simulated annealing protocol using the program X-PLOR²⁸. For the structure calculations, the protein was held fixed with the exception of those residues that form the ligand-binding site. NMR and structural statistics are given in Supplementary Tables 1–5.

Compound preparation and affinity. ABT-737 was synthesized as illustrated in Supplementary Fig. 1. *K_i* values were determined using competitive fluorescence polarization assays as described previously²⁹, and were calculated using the

equation for the binding of two competing ligands to a protein³⁰ ($n = 3$, mean $\pm$ s.e.m.).

Cytochrome *c* release from isolated mitochondria. Mitochondria were isolated from the mouse cell line FL5.12, incubated with peptide or drug for 35 min in the presence or absence of activated Bid (caspase-8-cleaved Bid myristoylated in a reaction with myristoyl CoA and N-myristoyl transferase, Myr-Bid), and cytochrome *c* release was quantified as previously described²².

Mammalian two hybrid assay. Full length Bcl-X_L and Bcl-X_S were cloned to the mammalian two hybrid system vectors pMD DNA BD and pVP16 AD, respectively (BD Bioscience). Bcl-X_L (3 μ g), Bcl-X_S (3 μ g) and pG5CAT indicator plasmids (3 μ g) were transfected into HeLa cells using Lipofectamine reagent (Invitrogen). After 24 h, cells were treated with compounds or DMSO for an additional 24 h. CAT activity was measured using CAT ELISA kit (Roche Applied Science). Inhibition was normalized to the DMSO control.

Confocal microscopy. Low-passage HeLa cells were transfected with Bcl-X_L and GFP-Bcl-G_s at a ratio of 10:1 using Lipofectamine plus reagent (Invitrogen). Time-lapsed imaging was performed using a laser-scanning confocal microscope MRC 1024-MP (BioRad) equipped with thermstage (Warner Instruments). The fluorescence of mitochondrial green fluorescent protein (GFP)-Bcl-G_s was excited with the 488-nm laser line and the Mitotracker red CMXRos (Molecular Probes) was excited with the 568-nm laser line. Confocal images were acquired 10 min after addition of the compound.

Cell culture of tumour cell lines. Cells were cultured with compound for 48 h in RPMI 1640 medium supplemented with 10% FBS, 1% sodium pyruvate, 4.5 g l⁻¹ glucose, and 1% antibiotic-antimycotic. Viability was determined by MTS assay (Promega). For potentiation studies, A549 NSCLC cells were incubated with paclitaxel alone in serum-free medium for 24 h followed by ABT-737 or the enantiomer for 48 h ($n = 3$, mean $\pm$ s.e.m.).

Primary patient tumour samples. Follicular lymphoma cells from anonymized donors were grown on a feeder layer of CD40 ligand-expressing 3T3 cells in Iscove's Modified Dulbecco's Medium supplemented with 10% Human AB serum, penicillin 100 U ml⁻¹, streptomycin 100 μ g ml⁻¹, 10 mM HEPES pH 7.4, and 2 mM L-glutamine for 40–44 h. For CLL specimens, peripheral blood was obtained from anonymized donors (samples a–g in Fig. 3c) or from patients who had given written informed consent (samples h–o in Fig. 3c). CLL cells were purified using Ficoll gradient and negative immunomagnetic selection, and were incubated in the same medium as the follicular lymphoma cells with the addition of transferrin (50 μ g ml⁻¹) and insulin (5 μ g ml⁻¹) for 48 h (samples a–g in Fig. 3c) or in RPMI 1640 medium, supplemented with 10%

FBS (fetal bovine serum) and penicillin/streptomycin and L-glutamine for 24 h (samples h–o in Fig. 3c). Cell death was quantified by fluorescence-activated cell sorting (FACS) analysis of annexin V-positive cells.

NCI-H146 and NCI-H1963 xenograft tumour studies. *Scid* (NCI-H146) or *scid-bg* (NCI-H1963) mice (Charles River Laboratories) were inoculated subcutaneously (s.c.) with 5×10^6 low-passage tumour cells mixed with Matrigel (BD Biosciences). Thirty-one days (NCI-H146) or thirty-two days (NCI-H1963) after inoculation, tumours were size-matched with an average volume of 225–230 mm³ with therapy initiated the following day. ABT-737 is not orally bioavailable in mice and was given every day (q.d.) intraperitoneally (i.p.) for 21 days ($n = 9$ –10 mice per group). Tumour volume was estimated by calliper measurements ($V = L \times W^2/2$). Complete regression (cure) was defined as absence of measurable tumour at the end of the experiment. For i.p. administration, 1 g ml⁻¹ stock solution of ABT-737 in DMSO was added to a mixture of 30% propylene glycol, 5% Tween 80, 65% D5W (5% dextrose in water), pH 4–5 (final concentration of DMSO $\leq 1\%$).

Immunohistochemistry. Mice with established H146 tumours were given a single i.p. injection of ABT-737 at 100 mg kg⁻¹ and tumours were collected at various times for immunohistochemical staining for activated caspase-3 (rabbit anti-active caspase-3 at 1:400, BD Pharmingen). Caspase-positive cells were scored at 400 $\times$ magnification. The average number of positive cells per 0.0625 mm² area was determined from three separate fields in each of three independent tumour samples.

Received 10 December 2004; accepted 31 March 2005

Published online 15 May 2005.

- Daniel, N. N. & Korsmeyer, S. J. Cell death: critical control points. *Cell* **116**, 205–219 (2004).
- Kirkin, V., Joos, S. & Zörnig, M. The role of Bcl-2 family members in tumorigenesis. *Biochim. Biophys. Acta* **1644**, 229–249 (2004).
- Amundson, S. A. *et al.* An informatics approach identifying markers of chemosensitivity in human cancer cell lines. *Cancer Res.* **60**, 6101–6110 (2000).
- Reed, J. C. Promise and problems of Bcl-2 antisense therapy. *J. Natl Cancer Inst.* **89**, 988–990 (1997).
- Letai, A., Sorcinelli, M. D., Beard, C. & Korsmeyer, S. J. Antiapoptotic Bcl-2 is required for maintenance of a model leukemia. *Cancer Cell* **6**, 241–249 (2004).
- Klasa, R. J., Gillum, A. M., Klem, R. E. & Frankel, S. R. Oblimersen Bcl-2 antisense: facilitating apoptosis in anticancer treatment. *Antisense Nucleic Acid Drug Dev.* **12**, 193–213 (2002).
- Kutzi, O. *et al.* Development of a potent Bcl-XL antagonist based on α -helix mimicry. *J. Am. Chem. Soc.* **124**, 11838–11839 (2002).
- Tzung, S.-P. *et al.* Antimycin A mimics a cell-death-inducing Bcl-2 homology domain 3. *Nature Cell Biol.* **3**, 183–191 (2001).
- Becattini, B. *et al.* Rational design and real time, in-cell detection of the proapoptotic activity of a novel compound targeting Bcl-X_L. *Chem. Biol.* **11**, 389–395 (2004).
- Kitada, S. *et al.* Discovery, characterization, and structure–activity relationships studies of proapoptotic polyphenols targeting B-cell lymphocyte/leukemia-2 proteins. *J. Med. Chem.* **46**, 4259–4264 (2003).
- Wang, J.-L. *et al.* Structure-based discovery of an organic compound that binds Bcl-2 protein and induces apoptosis of tumor cells. *Proc. Natl Acad. Sci. USA* **97**, 7124–7129 (2000).
- Degterev, A. *et al.* Identification of small-molecule inhibitors of interaction between the BH3 domain and Bcl-X_L. *Nature Cell Biol.* **3**, 173–182 (2001).
- Enyedy, I. J. *et al.* Discovery of small-molecule inhibitors of Bcl-2 through structure-based computer screening. *J. Med. Chem.* **44**, 4313–4324 (2001).
- Walensky, L. D. *et al.* Activation of apoptosis *in vivo* by a hydrocarbon-stapled BH3 helix. *Science* **305**, 1466–1470 (2004).
- Baell, J. B. & Huang, D. C. S. Prospects for targeting the Bcl-2 family of proteins to develop novel cytotoxic drugs. *Biochem. Pharm.* **64**, 851–863 (2002).
- Petros, A. M., Olejniczak, E. T. & Fesik, S. W. Structural biology of the Bcl-2 family of proteins. *Biochim. Biophys. Acta* **1644**, 83–94 (2004).
- Kelekar, A. & Thompson, C. B. Bcl-2-family proteins: the role of the BH3 domain in apoptosis. *Trends Cell Biol.* **8**, 324–330 (1998).
- Huang, D. C. & Strasser, A. BH3-only proteins—essential initiators of apoptotic cell death. *Cell* **103**, 839–842 (2000).
- Sattler, M. *et al.* Structure of Bcl-X_L-Bak peptide complex: Recognition between regulators of apoptosis. *Science* **275**, 983–986 (1997).
- Shuker, S. B., Hajduk, P. J., Meadows, R. P. & Fesik, S. W. Discovering high-affinity ligands for proteins: SAR by NMR. *Science* **274**, 1531–1534 (1996).
- Mao, H. *et al.* Rational design of difluorinated analogues with reduced affinity for human serum albumin. *J. Am. Chem. Soc.* **123**, 10429–10435 (2001).
- Letai, A. *et al.* Distinct BH3 domains either sensitize or activate mitochondrial apoptosis, serving as prototype cancer therapeutics. *Cancer Cell* **2**, 183–192 (2002).
- Wei, M. C. *et al.* Proapoptotic BAX and BAK: A requisite gateway to mitochondrial dysfunction and death. *Science* **292**, 727–730 (2001).
- Fearon, E. R. *et al.* Karyoplasmic interaction selection strategy: A general strategy to detect protein–protein interactions in mammalian cells. *Proc. Natl Acad. Sci. USA* **89**, 7958–7962 (1992).
- Tsujimoto, Y., Finger, L. R., Yunis, J., Nowell, P. C. & Croce, C. M. Cloning of the chromosome breakpoint of neoplastic B cells with the t(14;18) chromosome translocation. *Science* **226**, 1097–1099 (1984).
- Petros, A. M. *et al.* Solution structure of the antiapoptotic protein bcl-2. *Proc. Natl Acad. Sci. USA* **98**, 3012–3017 (2001).
- Mao, H., Gunasekera, A. H. & Fesik, S. W. Expression, refolding, and isotopic labeling of human serum albumin domains for NMR spectroscopy. *Protein Express. Purif.* **20**, 492–499 (2000).
- Brunger, A. T. *X-PLOR Version 3.1* (Yale Univ. Press, New Haven/London, 1992).
- Zhang, H., Nimmer, P., Rosenberg, S. H., Ng, S.-C. & Joseph, M. Development of a high-throughput fluorescence polarization assay for Bcl-X_L. *Anal. Biochem.* **307**, 70–75 (2002).
- Wang, Z.-X. An exact mathematical expression for describing competitive binding of two different ligands to a protein molecule. *FEBS Lett.* **360**, 111–114 (1995).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank T. J. Kipps, L. Rassenti, J. Gribben and L. Vallat for CLL Research Consortium samples, E. Monosov for assistance with confocal imaging, C. Rudin and C. Hann for H345 xenograft studies, and S. Ackler for compound formulation.

Author Information Atomic coordinates have been deposited in the Brookhaven Protein Data Bank under the accession codes 1YSG (Fig. 1a), 1YSI (Fig. 1b), 1YSX (Fig. 1c), 1YSN (Supplementary Fig. 2a), and 1YSW (Supplementary Fig. 2b). Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare competing financial interests: details accompany the paper on www.nature.com/nature. Correspondence and requests for materials should be addressed to S.H.R. (saul.rosenberg@abbott.com) or S.W.F. (stephen.fesik@abbott.com).

LETTERS

SV40-encoded microRNAs regulate viral gene expression and reduce susceptibility to cytotoxic T cells

Christopher S. Sullivan¹, Adam T. Grundhoff¹, Satvir Tevethia², James M. Pipas³ & Don Ganem¹

MicroRNAs (miRNAs) are small (~22-nucleotide) RNAs that in lower organisms serve important regulatory roles in development and gene expression, typically by forming imperfect duplexes with target messenger RNAs¹. miRNAs have also been described in mammalian cells and in infections with Epstein–Barr virus (EBV), but the function of most of them is unknown. Although one EBV miRNA probably altered the processing of a viral mRNA², the regulatory significance of this event is uncertain, because other transcripts exist that can supply the targeted function³. Here we report the identification of miRNAs encoded by simian virus 40 (SV40) and define their functional significance for viral infection. SVmiRNAs accumulate at late times in infection, are perfectly complementary to early viral mRNAs, and target those mRNAs for cleavage. This reduces the expression of viral T antigens but does not reduce the yield of infectious virus relative to that generated by a mutant lacking SVmiRNAs. However, wild-type SV40-infected cells are less sensitive than the mutant to lysis by cytotoxic T cells, and trigger less cytokine production by such cells. Thus, viral evolution has taken advantage of the miRNA pathway to

generate effectors that enhance the probability of successful infection.

To search for virally encoded miRNAs, we developed a computer program to predict likely miRNA precursors (pre-miRNAs) and performed a screen *in silico* for these RNAs in SV40. After searching the entire viral genome in both orientations, we found only two candidates, one each in the early and late orientations (Supplementary Fig. 1). The early-polarity pre-miRNA was predicted at nucleotide positions 29–5,201. However, northern blotting of RNA from infected cells with a probe spanning this region detected no small RNAs of appropriate size (data not shown). The predicted late-polarity pre-miRNA maps to the untranslated region 3' to the polyadenylation cleavage site in the late pre-mRNA (Fig. 1a). An RNA in this position has not previously been reported; however, this region is just 5' to (and overlaps by 20 nucleotides (nt)) a ~62-nt transcript of unknown function called SAS (SV40-associated small RNA), initially discovered in 1980 (ref. 4). Several different oligonucleotides complementary to this region were used to probe SV40 total RNA (Fig. 1b). A probe that spans most of the computer-predicted

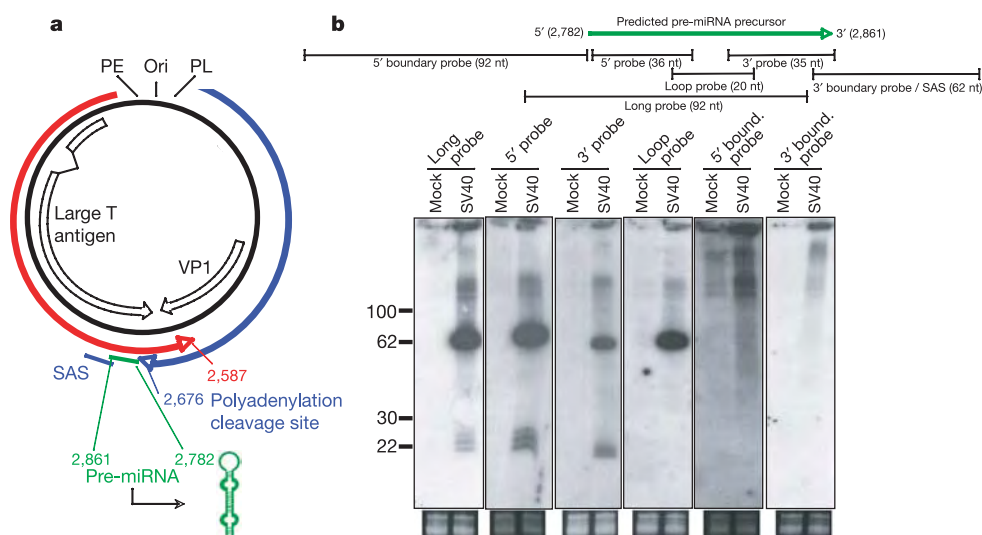


Figure 1 | SV40 encoded miRNAs. **a**, Map showing location of early transcripts (red), late transcripts (blue), SVpre-miRNA (green) and selected viral coding regions (boxes). PE, early promoter; PL, late promoter. **b**, Northern blot analysis for SV40 miRNAs in SV40-infected and

mock-infected BSC40 cells at 70 h after infection. The migration of a DNA oligonucleotide marker ladder is indicated at the left of the gels. Bound., boundary; ori, origin.

¹Howard Hughes Medical Institute and Departments of Microbiology and Medicine, G. W. Hooper Foundation, University of California, San Francisco, California 94143-0552, USA. ²Department of Microbiology and Immunology, The Pennsylvania State College of Medicine, Hershey, Pennsylvania 17033, USA. ³Department of Biological Sciences, University of Pittsburgh, Pittsburgh, Pennsylvania 15260, USA.

hairpin (long probe) shows a predominant ~60-nt band and several smaller ~22-nt RNAs; these are consistent with a pre-miRNA and processed miRNAs. As expected, probes that span either side of the predicted loop structure (5' or 3' probes) also detect the pre-miRNA; however, both probes also detect ~22-mers. For most miRNAs, only one strand of the hairpin pre-miRNA is energetically favoured to enter the RNA-induced silencing complex (RISC) and accumulate as a stable miRNA⁵. The fact that both strands are readily detectable as 22-mers indicates that the SV40 pre-miRNA belongs to a subset of pre-miRNAs in which both strands are incorporated into RISC. As a control, a probe designed for the loop portion (loop probe) of the predicted hairpin structure identifies only the pre-miRNA 60-nt band and not the processed 22-mers. Finally, probes that recognize regions predominantly outside the predicted hairpin structure—the 5' boundary probe and the 3' boundary probe (which is completely homologous to SAS)—do not detect any bands (Fig. 1b). These data show that the 22-mers detected from this region have a discrete sequence composition consistent with their being miRNA derivatives of the pre-miRNA, and that they are not simply degradation products from larger RNAs. The results also indicate that the transcripts being detected here are not the previously reported SAS RNA (see below).

To map the SVpre-miRNAs and SVmiRNAs with more resolution,

RNase protection assays (RPAs) were used. Figure 2a illustrates our RNase mapping strategy. We size-selected either pre-miRNA (45–65 nt; left panel) or miRNAs (15–35 nt) from infected-cell RNA. The pre-miRNA was mapped with the three probes whose SV40 sequence content is indicated in the left panel (all probes contained additional, irrelevant sequences flanking the SV40 sequence). Of these probes, the middle probe was predicted to be fully protected, whereas the 5' and 3' probes were expected to be only partly protected. As shown in Fig. 2b, the specifically protected fragment of the middle probe (asterisk) was 30 nt long—exactly as predicted. The 5' and 3' probes protected bands of 19–23 and 21–24 nt, respectively (asterisks). This maps the 5' end of the pre-miRNA to SV40 nt 2,794, and the 3' end to nt 2,851. A similar strategy was used to map the two sets of processed miRNAs (Fig. 2a, right panel). The 5' probe specifically protected bands of 18–22 nt (Fig. 2c, asterisks); the 3' probe protected bands of 18 or 19 nt (Fig. 2c, asterisks). This allows an approximation of the extent of the two sets of miRNAs, summarized (in red and blue ink) on the structure of the pre-miRNA in Fig. 2d.

Several conclusions can be drawn from these studies (see Fig. 2d). First, the longest pre-miRNA structure mapped is 57 nt long, in good agreement with its size on the northern blot of Fig. 1, and its most stable predicted hairpin fold has a 3' OH overhang consistent with

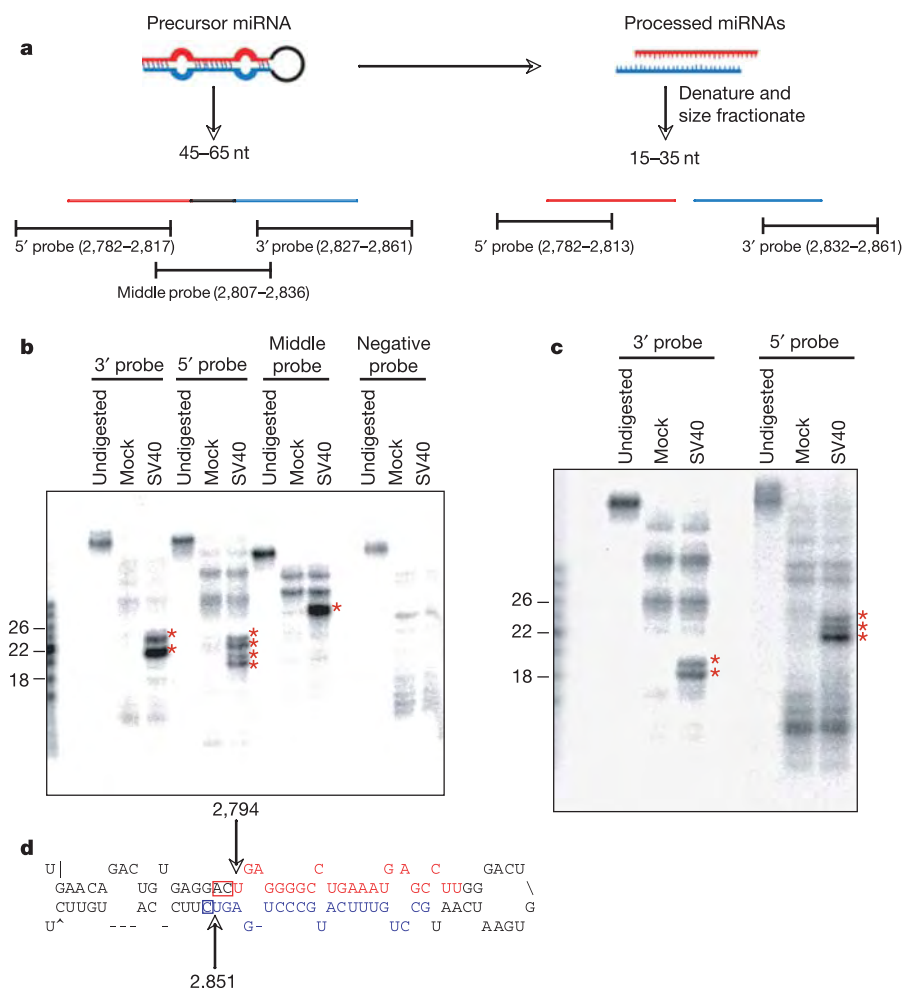


Figure 2 | Fine mapping of SV40 pre-miRNA and miRNAs. a, Diagram showing the probes and source of RNA used in the RPA studies. **b**, RPA mapping of SV40 pre-miRNA on size-fractionated ~45–65-mer RNA. **c**, RPA mapping of SV40 miRNA on size-fractionated 15–30-mer RNA. **d**, Predicted structure of SVpre-miRNA. miRNAs deduced from RPA and

northern blot results are indicated in red and blue, respectively. Arrows that correspond to the most abundant miRNAs indicate the nucleotide positions of the termini of the largest mapped pre-miRNA. Alternative termini determined from miRNA mapping studies are indicated with red and blue boxes.

Drosha processing^{6,7}. Second, an identical number of miRNAs are identified with the northern blot and RPA mapping: two miRNAs generated on the 3' side of the hairpin and three from the 5' side (Fig. 2c). Third, the ends of the predominant miRNAs map to within 1 nt of the mapped end of the SVpre-miRNA (Fig. 2b), which is consistent with processing from pre-miRNA to miRNA. Last, an RPA probe designed to detect SAS (negative probe) is also negative, which is consistent with the northern blot results of Fig. 1. Given the proximity of SVpre-miRNA and the reported position of SAS, and the fact that the SVpre-miRNA and SAS are similar in size (57 and 62 nt, respectively), we think it likely that the RNA originally identified as SAS was mis-mapped by about 40 nt, and is in fact the pre-miRNA we have identified.

We next determined the temporal expression pattern of the SVmiRNAs. Northern blot analysis shows that the SVpre-miRNA and the SVmiRNA are expressed late in infection, which is consistent with their processing from a late viral transcript (Fig. 3a). Because both sets of miRNAs are predicted to be perfectly complementary to the early mRNAs, we proposed that the SVmiRNAs might direct the cleavage of the early mRNAs by means of the RNA-mediated interference (RNAi) machinery, just as short interfering RNAs would do. To test this model, we prepared total and poly(A)⁺ RNA isolated from cells infected with SV40 at various times after infection, and northern blotted with strand-specific probes complementary to regions 3' or 5' to the predicted cleavage sites in early mRNA. We observed a heterogeneous ~260-nt species that was present in polyadenylated and total RNA (Fig. 3b). This RNA was detected only with the 3' probe (Fig. 3b) and corresponds in size to the 3' product of cleavage of early mRNA, including its poly(A) tail (which accounts for the observed heterogeneity of its electrophoretic mobility). We do not detect the stable accumulation of the intact 5' product of this predicted cleavage (Fig. 3b, lower panel); this is in accord with earlier findings that the 3' product of RNA cleaved by the RNAi machinery is more stable than the 5' product⁸. (The ~1.9-kilobase band migrating faster than the early mRNA is too

small to be the 5' product expected from the SVmiRNA-mediated cleavage. However, it might be a stable degradation product of that species.)

We next conducted an RPA analysis on poly(A)⁺ RNA from infected cells to map the 5' end of the 260-nt 3' cleavage product more precisely (Fig. 3c). We observed three predominant protected fragments, one corresponding to the full-length (uncleaved) early transcripts and two others that map approximately to nt 2,844 and 2,802—precisely within the regions homologous to each set of miRNAs. This result indicates that both sets of miRNAs might actively direct the cleavage of their complements on the early mRNAs. To our knowledge, this is the first example of a single pre-miRNA encoding miRNAs from both arms of the hairpin that are active against the same target.

To explore the role of SVmiRNAs in SV40 biology, a mutant, SM (for SV40 miRNA mutant), was constructed by mutagenizing carefully selected bases in the region of the predicted pre-miRNA (Supplementary Fig. 2). The lesions were chosen to disrupt the hairpin structure of the pre-miRNA on the late strand but leave intact the amino-acid coding potential of T antigen on the early strand (Supplementary Fig. 2). We observed that, unlike the wild type, this mutant does not generate the early mRNA cleavage fragments and possesses higher levels of intact early RNA at all time points tested (Fig. 4a). This proves that the early mRNA fragments observed in Fig. 3 are indeed dependent on the SVmiRNAs for their production, and underscores that these miRNAs are likely to serve a negative regulatory role *in vivo*. Correspondingly, the production of large and small T antigens by SM was enhanced relative to wild-type (Fig. 4b). Importantly, though miRNA action downregulated antigen expression, it was not associated with a reduced generation of infectious virus, as judged by one-step growth curves (Fig. 4c) and by analysis of multiple cycles of viral replication (data not shown). Consistent with this was the observation that levels of late viral structural proteins were also unaffected (Fig. 4b).

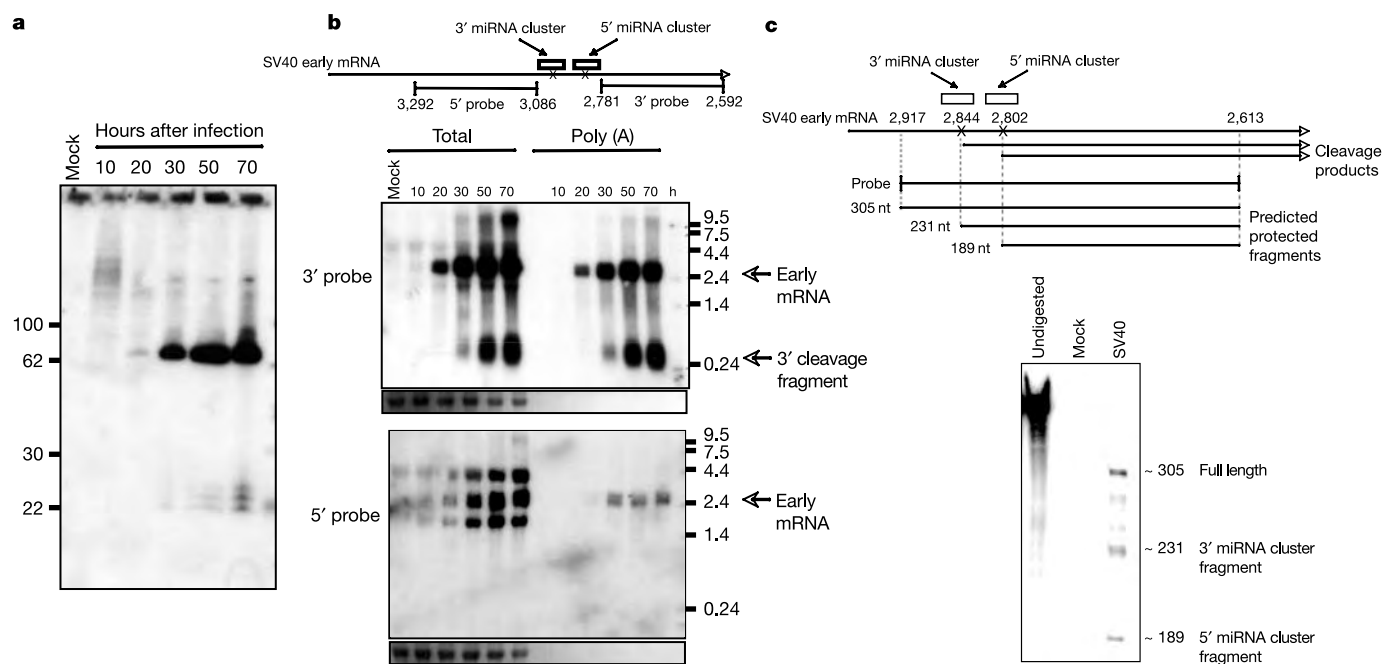


Figure 3 | A role for SV40miRNAs in cleavage of early mRNAs. **a**, Northern blot for SVmiRNAs at various time points after infection. **b**, Northern blot for early mRNAs of total and polyadenylate-enriched RNA (poly(A)). Predicted miRNA-directed cleavage sites for each miRNA cluster are indicated with crosses. Top, probe recognizes sequences 3' to predicted

cleavage site; bottom, probe recognizes sequences 5' to predicted cleavage site. Below each panel, as a loading control, is shown methylene blue staining of 28S ribosomal RNA. A marker ladder in kilobases is shown to the right of each gel. **c**, RPA mapping of poly(A)-enriched RNA from 70 h after infection.

What is the functional significance of this T-antigen downregulation? The observation that yields of infectious virus are unaffected in the SM mutant means that the excess T antigen produced in the absence of the miRNAs serves no essential replicative function. But apart from its roles in viral replication, T antigen is also a target of the cytotoxic T lymphocyte (CTL) response⁹. This led us to consider the possibility that one function of the miRNAs might be the promotion of CTL evasion. Accordingly, we examined the susceptibility of wild-type SV40-infected and SM-infected cells to lysis by CTL clones specific for SV40 T antigen. As shown in Fig. 4d, under conditions that mimic those that would be found in infected tissues *in vivo* during primary infection (low multiplicity of infection (MOI), low CTL:target ratios), wild-type-infected cells have a significantly lower susceptibility to CTL-mediated lysis than their SM-infected counterparts. This was true for two independent CTL clones recognizing different epitopes in T antigen, and true at several CTL:target ratios (Fig. 4d, Supplementary Fig. 3a). In addition, release of

interferon- γ was also diminished when CTLs encountered infected cells expressing the SVmiRNAs (Supplementary Fig. 3b).

SVmiRNAs are, to our knowledge, the first miRNAs shown to have a function in virus biology. By downregulating the accumulation of unnecessary T antigen, the SVmiRNAs reduce CTL susceptibility and local cytokine release. Thus, although this downregulation is dispensable for viral growth in culture, it is likely to be of considerable importance *in vivo*. This is consistent with the fact that the predicted hairpin structure for the pre-miRNA is not only found in all SV40 isolates but also conserved in other primate polyomaviruses, including BKV, JCV and SA12 (Supplementary Fig. 4). The existence of these autoregulatory miRNAs indicates that viruses can use the host RNAi machinery, which is often speculated to have evolved as an antiviral mechanism, to generate small RNAs that serve their own purposes—the latest chapter in the long cat-and-mouse game known to virologists as host–virus coevolution.

METHODS

Computational methods. An algorithm, VirMir, was written with an easy user interface to identify likely pre-miRNAs in small genomes (less than 300 kilobases). Windows of 100 nt were tiled across the viral genomes in both orientations at 10-nt intervals. These windows were submitted to RNAfold¹⁰ and the hairpin structures obtained were each scored independently. Individual hairpins were scored as follows: +1 for each paired nucleotide, -0.5 for each unpaired nucleotide in a symmetric bulge, -2 per nucleotide for symmetric bulges greater than 2 nt and for asymmetric bulges, -1 per nucleotide for each nucleotide greater than 5 in the terminal loop. This score was multiplied by the free energy, and the resulting best-scoring individual hairpin was plotted as the overall score for each window.

RNA folding predictions shown in Fig. 2 and Supplementary Fig. 4 were conducted with mFold¹¹.

SM virus construction and viral assays. To create SM, PCR was used to generate two fragments that were inserted into pSVB3 (ref. 12) through a three-way ligation with *NheI*, *AgeI* and *BstXI*. The primer sets used were 5'-TATCGTCC ATTCCGACAG-3' and 5'-TGGTCACGAGACCGGTATTGATTCTCAGTCGC AGGGTTCGTTTCAAGCTCCGAGTCTTCGAGTCTGTTTCATGATCATAA TCAGCCA-3', and 5'-ACTGCAACAATGGCCTG-3' and 5'-GAATCAATA CCGGTCTCATGACCAGAATCCTCCATATTCTTCTCCCACCATCTTCAT TTTTA-3'. SM was transfected into BSC40 cells and the resulting virus was plaque-purified and amplified as described previously¹³. The one-step growth curve shown in Fig. 4 is a representative plot of multiple experiments conducted at a high MOI of five plaque-forming units per cell.

Western blot, RNA isolation, northern blot and RPA analysis. Protein was harvested in NETN400 (20 mM Tris pH 8.0, 400 mM NaCl, 1 mM EDTA, 0.5% NP40) and probed with an antibody (pAb4119) recognizing the amino terminus of large and small T antigens. Total RNA was harvested with RNeasy and probes were generated with radioactive probes transcribed *in vitro* (Figs 2, 3b, 3c and 4b) or radiolabelled antisense oligonucleotides (Figs 1 and 3a). Where indicated, signal intensities were quantified on a Storm 860 PhosphorImager (Molecular Dynamics).

In vitro cytotoxicity assay. TC-7/Db cells, a continuous line of monkey kidney cells expressing mouse Db MHC class I antigen, were infected with wild-type SV40 or SM at a MOI of 1 for 2 h at 37°C. The unabsorbed viruses were neutralized with anti-SV40 sera, cells were washed and incubated in fresh medium for 40 h; after this the cells were harvested, labelled with ⁵¹Cr and reacted for 4 h with Db-restricted murine CTL clone TCR-V(V) (specific for Tag epitope V, residues 489–497). The supernatant was counted in a gamma counter and results are expressed as percentage specific lysis, as described previously^{9,14}. The interferon- γ release assay was performed as described previously⁹.

Received 17 February; accepted 1 April 2005.

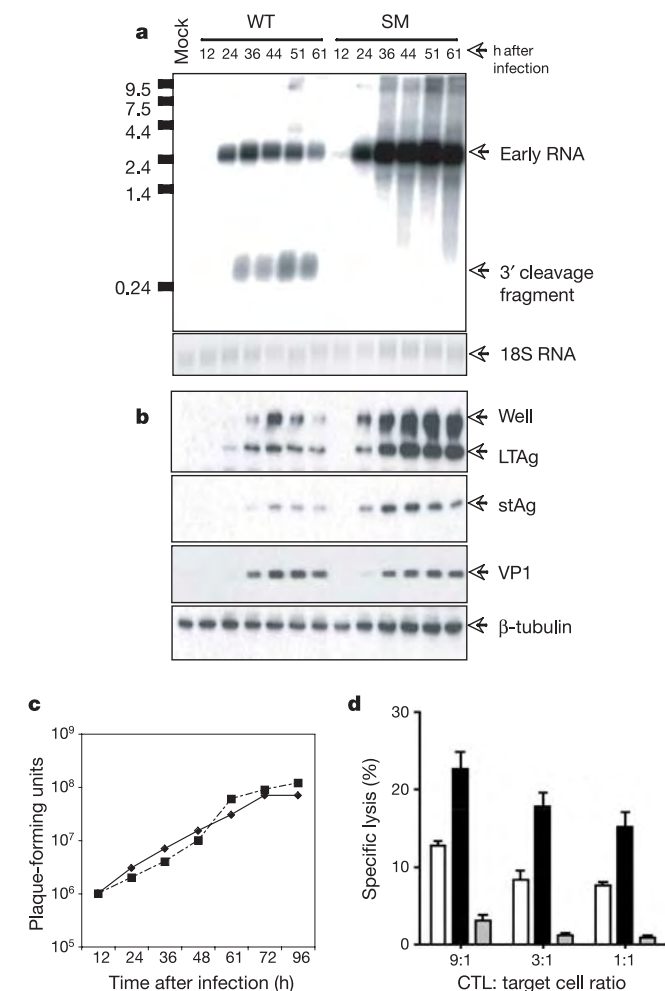


Figure 4 | Phenotype of the SM mutant. **a**, Northern blot of early RNAs conducted on total RNA. WT, wild-type SV40. Marker ladder in kilobases. **b**, Immunoblot showing protein levels of large T antigen (LTA), small t antigen (stAg), structural protein VP1 and β -tubulin as a loading control. **c**, Representative one-step growth curve of wild-type (diamonds) versus SM (squares). **d**, Simian TC7/Db cells (permissive for SV40 infection) that express the appropriate murine MHC-class I allele to allow antigen recognition were infected with wild-type SV40 (white bars) or SM (black bars) at a MOI of 1, labelled with ⁵¹Cr, then incubated with CTL clone TCR-V(V) specific for T-antigen residues 489–497. Grey bars, uninfected. Results are means and s.d. for triplicate samples.

- Ambros, V. The functions of animal microRNAs. *Nature* **431**, 350–355 (2004).
- Pfeffer, S. *et al.* Identification of virus-encoded microRNAs. *Science* **304**, 734–736 (2004).
- Furnari, F. B., Adams, M. D. & Pagano, J. S. Unconventional processing of the 3' termini of the Epstein–Barr virus DNA polymerase mRNA. *Proc. Natl Acad. Sci. USA* **90**, 378–382 (1993).
- Alwine, J. C. & Khoury, G. Simian virus 40-associated small RNA: mapping on the simian virus 40 genome and characterization of its synthesis. *J. Virol.* **36**, 701–708 (1980).
- Schwarz, D. S. *et al.* Asymmetry in the assembly of the RNAi enzyme complex. *Cell* **115**, 199–208 (2003).

6. Lee, Y. *et al.* The nuclear RNase III Drosha initiates microRNA processing. *Nature* **425**, 415–419 (2003).
7. Zeng, Y., Yi, R. & Cullen, B. R. MicroRNAs and small interfering RNAs can inhibit mRNA expression by similar mechanisms. *Proc. Natl Acad. Sci. USA* **100**, 9779–9784 (2003).
8. Hohen, T., Amarzguoui, M., Wiiger, M. T., Babaie, E. & Prydz, H. Positional effects of short interfering RNAs targeting the human coagulation trigger tissue factor. *Nucleic Acids Res.* **30**, 1757–1766 (2002).
9. Mylin, L. M. *et al.* Quantitation of CD8⁺T-lymphocyte responses to multiple epitopes from simian virus 40 (SV40) large T antigen in C57BL/6 mice immunized with SV40, SV40 T-antigen-transformed cells, or vaccinia virus recombinants expressing full-length T antigen or epitope minigenes. *J. Virol.* **74**, 6922–6934 (2000).
10. Hofacker, I. L., Fontana, W., Stadler, P. F., Bonhoeffer, M. & Schuster, P. Fast folding and comparison of RNA secondary structures. *Monatshefte Chemie* **125**, 167–188 (1994).
11. Zuker, M. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* **31**, 3406–3415 (2003).
12. Peden, K. W., Pipas, J. M., Pearson-White, S. & Nathans, D. Isolation of mutants of an animal virus in bacteria. *Science* **209**, 1392–1396 (1980).
13. Tremblay, J. D., Sachsenmeier, K. F. & Pipas, J. M. Propagation of wild-type and mutant SV40. *Methods Mol. Biol.* **165**, 1–7 (2001).
14. Bates, M. P., Jennings, S. R., Tanaka, Y., Tevethia, M. J. & Tevethia, S. S. Recognition of simian virus 40 T antigen synthesized during viral lytic cycle in

monkey kidney cells expressing mouse H-2Kb- and H-2Db-transfected genes by SV40-specific cytotoxic T lymphocytes leads to the abrogation of virus lytic cycle. *Virology* **162**, 197–205 (1988).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank M. McManus, J. Ziegelbauer and J. Bechtel for discussions; P. Othol and M. Epler for their assistance in CTL assays; and B. Garcea for the gift of VP1 antibody. This work was supported by the Howard Hughes Medical Institute. S.T. and J.M.P. are supported by NIH grants.

Author Contributions C.S.S. conceived the project and carried out all experiments in Figs 1–4c. D.G. directed and supervised the experimental work and interpretation. A.T.G. developed and applied the computational algorithm for miRNA detection, and assisted in the design of the SM mutant. J.M.P. assisted with project planning, provided viral strains and participated in data review. S.T. designed and carried out the CTL lysis experiments. All authors participated in the drafting of the manuscript.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to D.G. (ganem@cgl.ucsf.edu).

Insights into E3 ligase activity revealed by a SUMO-RanGAP1-Ubc9-Nup358 complex

David Reverter¹ & Christopher D. Lima¹

SUMO-1 (for small ubiquitin-related modifier) belongs to the ubiquitin (Ub) and ubiquitin-like (Ubl) protein family. SUMO conjugation occurs on specific lysine residues within protein targets, regulating pathways involved in differentiation, apoptosis, the cell cycle and responses to stress by altering protein function through changes in activity or cellular localization or by protecting substrates from ubiquitination^{1,2}. Ub/Ubl conjugation occurs in sequential steps and requires the concerted action of E2 conjugating proteins and E3 ligases^{1,2}. In addition to being a SUMO E3, the nucleoporin Nup358/RanBP2 localizes SUMO-conjugated RanGAP1 to the cytoplasmic face of the nuclear pore complex by means of interactions in a complex that also includes Ubc9, the SUMO E2 conjugating protein^{3–6}. Here we describe the 3.0-Å crystal structure of a four-protein complex of Ubc9, a Nup358/RanBP2 E3 ligase domain (IR1-M) and SUMO-1 conjugated to the carboxy-terminal domain of RanGAP1. Structural insights, combined with biochemical and kinetic data obtained with additional substrates, support a model in which Nup358/RanBP2 acts as an E3 by binding both SUMO and Ubc9 to position the SUMO-E2-thioester in an optimal orientation to enhance conjugation.

Ub/Ubls are activated by E1 and transferred to E2 to form E2-Ub/Ubl-thioesters. Although competent for Ub/Ubl ligation to lysine ε-amino groups, E2s generally require E3 ligases to recognize substrate lysine residues specifically. Most E3s belong to either RING or HECT families, and whereas RING E3s recruit substrate and bind the E2-Ub/Ubl through a zinc domain to promote conjugation to lysine residues⁷, HECT E3s recruit E2-Ub/Ubl to generate E3-Ub/Ubl-thioesters for conjugation⁸. The SUMO E2 can directly bind the consensus sequence Ψ-K-X-D/E where Ψ is a hydrophobic amino acid, K is the substrate lysine, X is any amino acid and D or E is acidic⁹, although several SUMO E3s facilitate conjugation *in vivo* and *in vitro*; they include RING-type E3s^{10,11}, Nup358/RanBP2 (ref. 12) and Pc2 (ref. 13). Nup358/RanBP2 and Pc2 seem unrelated to either RING or HECT E3 ligases.

One of the first discovered functions for SUMO-1 was its role in nucleocytoplasmic trafficking^{3–5}. SUMO conjugation localizes RanGAP1 to the nuclear pore complex (NPC) in a complex that includes the SUMO E2 Ubc9 and Nup358/RanBP2, a multi-domain 3,224-residue nucleoporin that also interacts with Ran and other nuclear transport factors^{3–6,14,15}. The SUMO-RanGAP1-Ubc9-Nup358/RanBP2 complex does not dissociate on entry into mitosis and NPC disassembly, but redistributes to kinetochores and contributes to the stability of kinetochore-microtubule interactions¹⁶. A roughly 30-kDa Nup358/RanBP2 fragment named IR1-M-IR2 binds Ubc9 and promotes SUMO E3 activity *in vitro* and *in vivo*^{12,17–19}, although domains flanking IR1-M-IR2 also contribute to SUMO-RanGAP1 interactions⁶. IR1-M-IR2 was parsed into three elements: IR1 (residues 2,633–2,685), M (residues 2,686–2,710) and

IR2 (residues 2,711–2,761). IR1 and IR2 were named for two internal sequence repeats. IR1-M-IR2, IR1-M and M-IR2 all promote SUMO-1 conjugation *in vitro*^{18,19}.

To determine the molecular details of this system, SUMO-1 was conjugated to RanGAP1, combined with Ubc9 and Nup358/RanBP2, purified by gel filtration, and crystallized. The model containing SUMO-1 (residues 20–97), RanGAP1 (residues 432–587), Ubc9 (residues 2–157) and Nup358/RanBP2 (residues 2,631–2,693) was refined to 3.0 Å ($R = 24.7$, $R_{\text{free}} = 29.0$; Table 1, Supplementary Table 1 and Supplementary Methods). The structure reveals Ubc9 as the central component of the quaternary complex, contacting SUMO, RanGAP1 and Nup358/RanBP2 (Fig. 1). SUMO contacts Ubc9 and Nup358/RanBP2, but contacts RanGAP1 only through the covalent bond between SUMO Gly 97 and Lys 524. Nup358/RanBP2 contacts SUMO and Ubc9 but does not contact RanGAP1.

The human SUMO-1-RanGAP1-Ubc9 complex is similar to our previously described complex between mouse RanGAP1 and human Ubc9 in that most interactions to the RanGAP1 Ψ-K-X-D/E SUMO motif (L-K-S-E) are preserved⁹ (Fig. 2). Despite covalent attachment to SUMO Gly 97 via an isopeptide bond, Lys 524 remains within a groove created by Ubc9 Asp 127, Pro 128, Ala 129 and Tyr 87. SUMO-1 Gln 29 and Arg 63 and the C-terminal tail (Gln 92, Gln 94, Thr 95, Gly 96 and Gly 97) contact Ubc9 helix B and Ubc9 active-site residues, respectively (Fig. 2), burying only 650 Å² of total surface area²⁰. For comparison, Sae1/Sae2-SUMO-1 and Senp2-SUMO-1 bury 1,650 and 1,800 Å², respectively²¹. Residues in the interface between SUMO-1 and Ubc9 are conserved among SUMO and Ubc9 family members. The small interface between SUMO and Ubc9 is consistent with the complexes being primed to dissociate after conjugation. It is also consistent with the specificities of Ubc9-SUMO being achieved by the SUMO E1 activating enzyme, which brings SUMO and Ubc9 together to form the thioester adduct²¹.

Several observations indicate that the Ub/Ubl-E2-thioester might adopt a similar configuration to that observed in our complex before conjugation. First, NMR chemical shift perturbations observed for a

Table 1 | X-ray statistics

Resolution (Å)	30–3.0
$R_{\text{work}}/R_{\text{free}}$	0.247/0.290
Number of atoms	
Protein	3,564
Water	28
B-factors	
Protein	90
Water	51
R.m.s. deviations	
Bond lengths (Å)	0.006
Bond angles (°)	1.2

¹Structural Biology Program, Sloan-Kettering Institute, New York, New York 10021, USA.

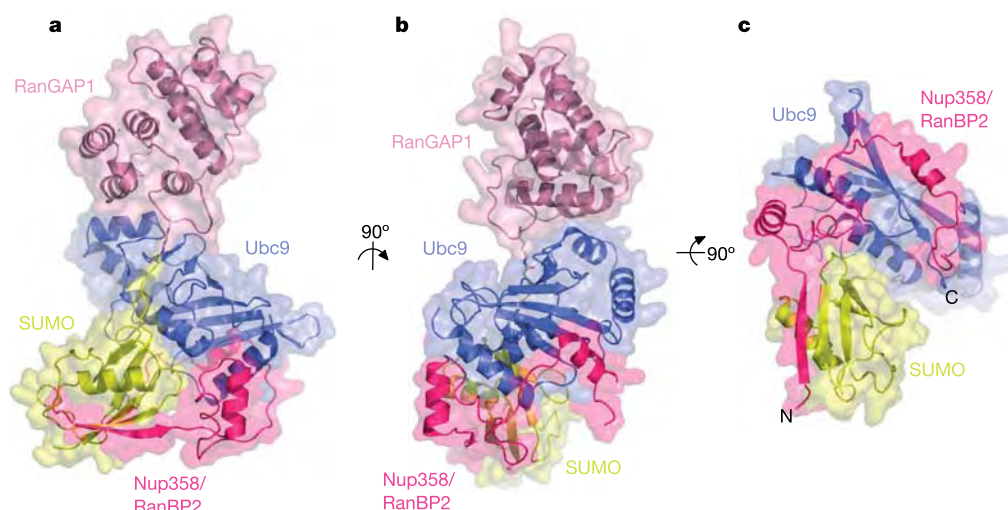


Figure 1 | Structure of SUMO-RanGAP1-Ubc9-Nup358/RanBP2 complex. **a**, Ribbon and transparent-surface representation for the complex between SUMO-1 (yellow), Ubc9 (blue), RanGAP1 (pink) and Nup358/RanBP2 (magenta). Each protein is labelled. The SUMO C-terminal glycine (Gly 97) and RanGAP1 Lys 524 are represented by solid bonds located near the image

centre. **b**, Orthogonal view of the complex. **c**, Orthogonal view of the complex to highlight the extended Nup358/RanBP2 structure. The N and C termini of Nup358/RanBP2 are indicated. Structural graphics were generated with Pymol (<http://pymol.sourceforge.net>).

steady-state Ub-Ubc1-thioester were consistent with contacts between Ub, E2 helix B and an E2 channel that coordinates the Ub Gly-Gly (ref. 22). Second, Ub/Ubls and E2s are structurally conserved. Third, a rotation about Ubc9 Cys 93 Chi1 brings the Cys sulphur atom to within 2 Å of the SUMO Gly 97 carbonyl carbon (Fig. 2). We previously indicated that E2s might coordinate the lysine near the labile E2-Ub/Ubl-thioester to promote chemistry⁹. The absence of other E2 catalytic residues led to recent studies that implicated the conserved E2 His-Pro-Asn amino acid motif (HPN) in catalysis, namely that Asn Nδ contributes to formation of an oxyanion hole that stabilizes the transition state²³. Our structure is consistent with this proposal in that as Asn 85 Nδ moves away from the Ubc9 main chain it approaches the C-terminal Gly 97 carbonyl oxygen, although definitive evidence for the catalytic role of Asn 85 is precluded by the resolution limits of our structure (3.0 Å) and because our complex includes a conjugated product rather than a E2-SUMO-thioester substrate complex.

Nup358/RanBP2 IR1-M adopts a non-globular and extended structure in the complex (Fig. 3). IR1-M elements were divided into motifs I–V based on interactions in the complex (Fig. 3a). Motif I forms an antiparallel β-strand with SUMO β-strand 2. In addition to main-chain contacts, motif I includes two acidic and five hydrophobic residues that contribute ionic and van der Waals contacts to SUMO-1 residues, respectively (Fig. 3c). The C-terminal end of IR1 α-helix A contacts the Ubc9 amino-terminal helix in addition to Pro 69, Pro 105 and Ala 106. Ubc9 Arg 8 bridges four main-chain carbonyl oxygen atoms, two from Ubc9 and two from Nup358/RanBP2 (Fig. 3d). Motif II forms a coil that packs hydrophobic residues into the interface formed by the Ubc9 N-terminal helix and residues between β1 and β2 of SUMO-1 (Fig. 3d). Motif III acidic residues bridge the Ubc9 N-terminal helix and contact Ubc9 Arg 13, Arg 17 and Lys 30 (Fig. 3e). IR1 then forms α-helix B, packing motif IV hydrophobic residues onto Ubc9 β1–β3 (Fig. 3f) before contacting Ubc9 β2, β3 and Phe 22 through Leu 2688–Tyr 2689–Leu 2690 in

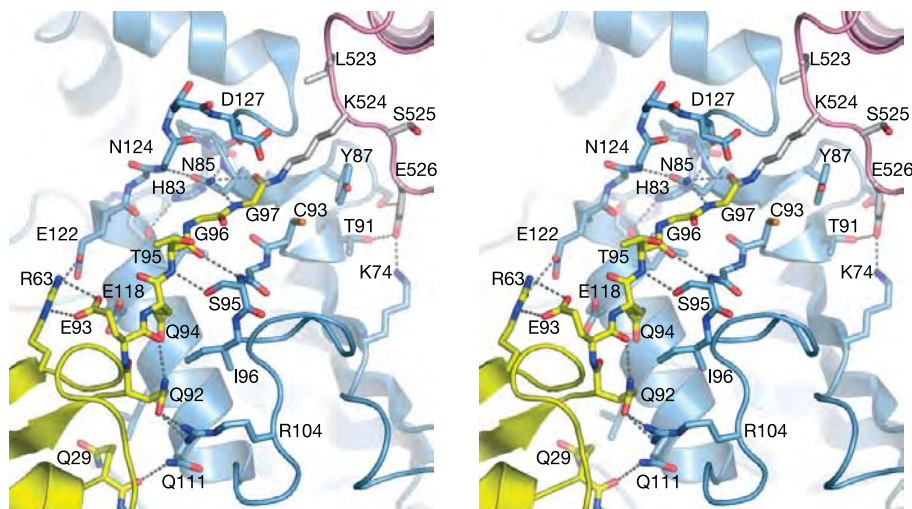


Figure 2 | E2 active site in complex with RanGAP1-SUMO-1. Stereo view of the E2 active site in complex with SUMO-1-RanGAP1 in ribbon and solid-bond representation. Residues are labelled, and hydrogen-bonding

interactions are indicated by dashed lines. SUMO-1, RanGAP1 and Ubc9 are coloured yellow, pink and blue as in Fig. 1.

motif V (Fig. 3g). The extended structure buries 1,460 Å² in the SUMO interface and 830 Å² in the Ubc9 interface. On the basis of our structure, new sequence alignments between IR1 and IR2 were used to generate IR1*, IR2* and IR1-M-IR2* (Fig. 3a).

RanGAP1 is easily conjugated and interacts strongly with Ubc9 through surfaces adjacent to the conjugation site⁹. It therefore remains unclear whether Nup358/RanBP2 E3 activity is required for RanGAP1 SUMO conjugation, or whether it is required merely to maintain a stable complex at the nuclear pore. The stable interactions observed between RanGAP1 and E2 and between E3, SUMO and E2, and the covalent interaction in RanGAP1–SUMO, indicated that our structure might represent a trapped product complex. To test this, conjugation assays with RanGAP1 under multiple-turnover conditions with and without Nup358/RanBP2 IR1* showed, in contrast to other substrates^{12,18,19}, that IR1* inhibited SUMO conjugation. SUMO–RanGAP1 accumulated to near-stoichiometric ratios of product, E2 and E3 (Fig. 4a). The addition of SUMO–RanGAP1 also inhibited SUMO conjugation to p53 in an IR1*-dependent manner (Fig. 4b). Again, stoichiometric ratios of SUMO-1–RanGAP1 to E2 seemed sufficient to sequester E2 and prevent further p53 conjugation. Partial disruption of the stable interface

between RanGAP1 and Ubc9 alleviated IR1*-dependent inhibition in multiple-turnover assays and rendered RanGAP1 a substrate for Nup358/RanBP2 E3 activity under single-turnover conditions (Supplementary Fig. 1).

To assess the importance of contacts between Nup358/RanBP2, SUMO and E2, assays were conducted at near-saturating substrate concentrations under multiple-turnover and single-turnover conditions to quantify rate velocities by using E3 constructs and substrates that included the p53 tetramerization domain, IκBα, and a peptide containing a SUMO consensus motif (Fig. 3a; see Methods). IR1* and IR1-M-IR2* enhanced conjugation rates up to 80-fold in multiple-turnover assays (Fig. 4c–e), and IR2* enhanced conjugation 4–16-fold. IR1T included only motifs I and II but catalysed E3 activity at a rate similar to that of IR2*, indicating that motifs III–V might be somewhat dispensable for activity. Deleting motif I (TIR1 and TIR2) from either IR1 or IR2 resulted in almost no E3 activity. Similar rate enhancements were observed under single-turnover conditions using isolated E2–SUMO-thioester (see Methods) (Fig. 4f–h).

Consistent with observations for motif I in our structure was the recent use of NMR to identify Nup358/RanBP2 IR1 amino acids

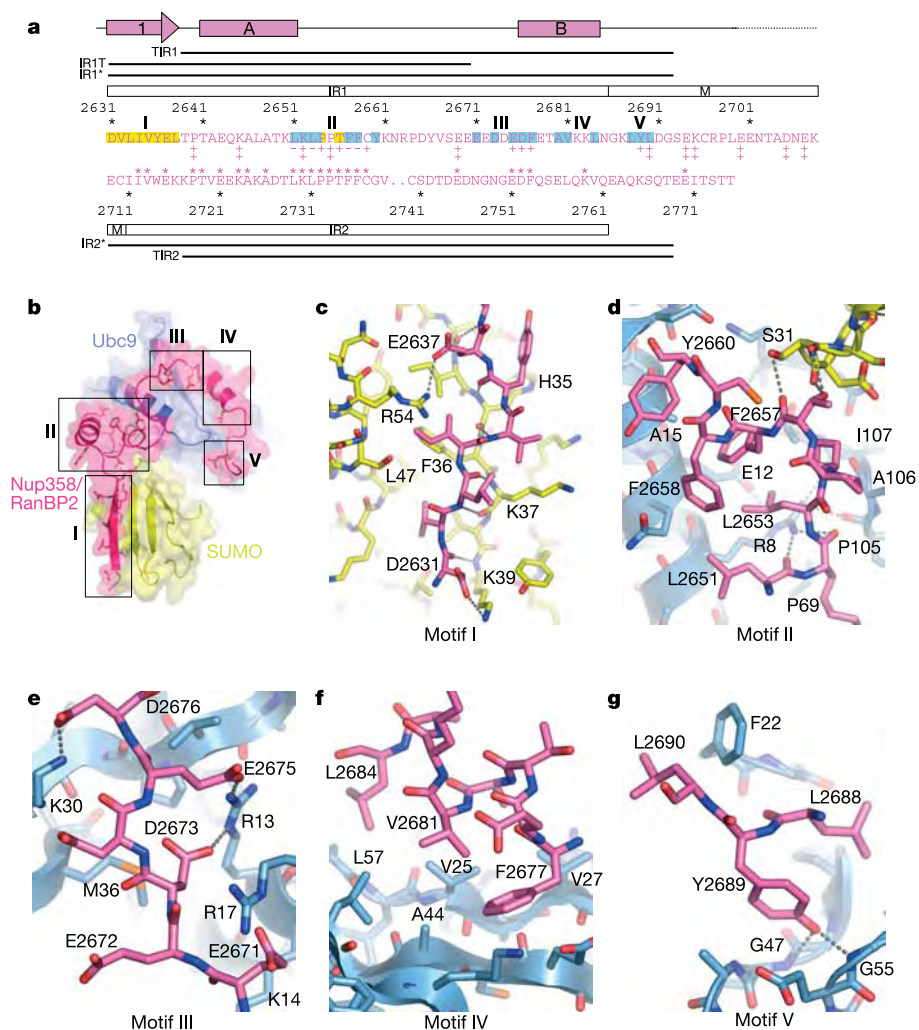


Figure 3 | Nup358/RanBP2 sequence alignment and E2-E3-SUMO-1 structure. **a**, IR1-M-IR2 elements. Secondary structure is shown above the alignment. Single-letter amino-acid code is coloured for SUMO (yellow) and Ubc9 (blue) contacts. Nup358/RanBP2 IR constructs are indicated by bars. Mutational analysis¹⁸ is shown below IR1-M: double dagger, no defect; plus sign, impaired activity; minus sign, no activity. Asterisks above IR2 indicate

identical amino-acid positions between IR1 and IR2. **b**, Nup358/RanBP2 motifs I–V (magenta), SUMO-1 (yellow) and Ubc9 (blue). **c**, Nup358/RanBP2 motif I. **d**, Nup358/RanBP2 motif II. **e**, Nup358/RanBP2 motif III. **f**, Nup358/RanBP2 motif IV. **g**, Nup358/RanBP2 motif V. Residues are labelled, and hydrogen-bonding interactions are indicated by dashed lines.

(motif I) that interact with SUMO; although E3 activity was not assessed, several mutations within motif I diminished binding to SUMO-1–RanGAP1 (Fig. 3c)²⁴. SUMO-1 interactions with Nup358/RanBP2 motif I might be recapitulated by other proteins that interact non-covalently with SUMO²⁴ and Smt3 (ref. 25). Mutational analysis revealed the importance of Nup358/RanBP2 motif II residues Leu 2651, Leu 2653, Phe 2657 and Phe 2658 in Ubc9 binding and E3 activity¹⁸, residues that contact Ubc9 in our structure (Fig. 3a, d). This and another study identified Ubc9 mutations that partly diminished both E3 activity and binding to Nup358/RanBP2

(refs 18, 19). Most detrimental Ubc9 mutations face Nup358/RanBP2 motifs IV and V in our structure. The latter study also indicated a possible mechanism for SUMO paralogue selection by various Nup358/RanBP2 IR1-M-IR2 domains¹⁹. To confirm that SUMO-2/SUMO-3–E2 is activated in an analogous manner to that proposed for SUMO-1–E2, we assayed SUMO-2 and SUMO-3 under single-turnover conditions. These data revealed rate enhancements for IR1-M-IR2* and IR1*, and a strict dependence on motif I for activity (Supplementary Fig. 2). If Nup358/RanBP2 binds E2–SUMO-thioester in a manner similar

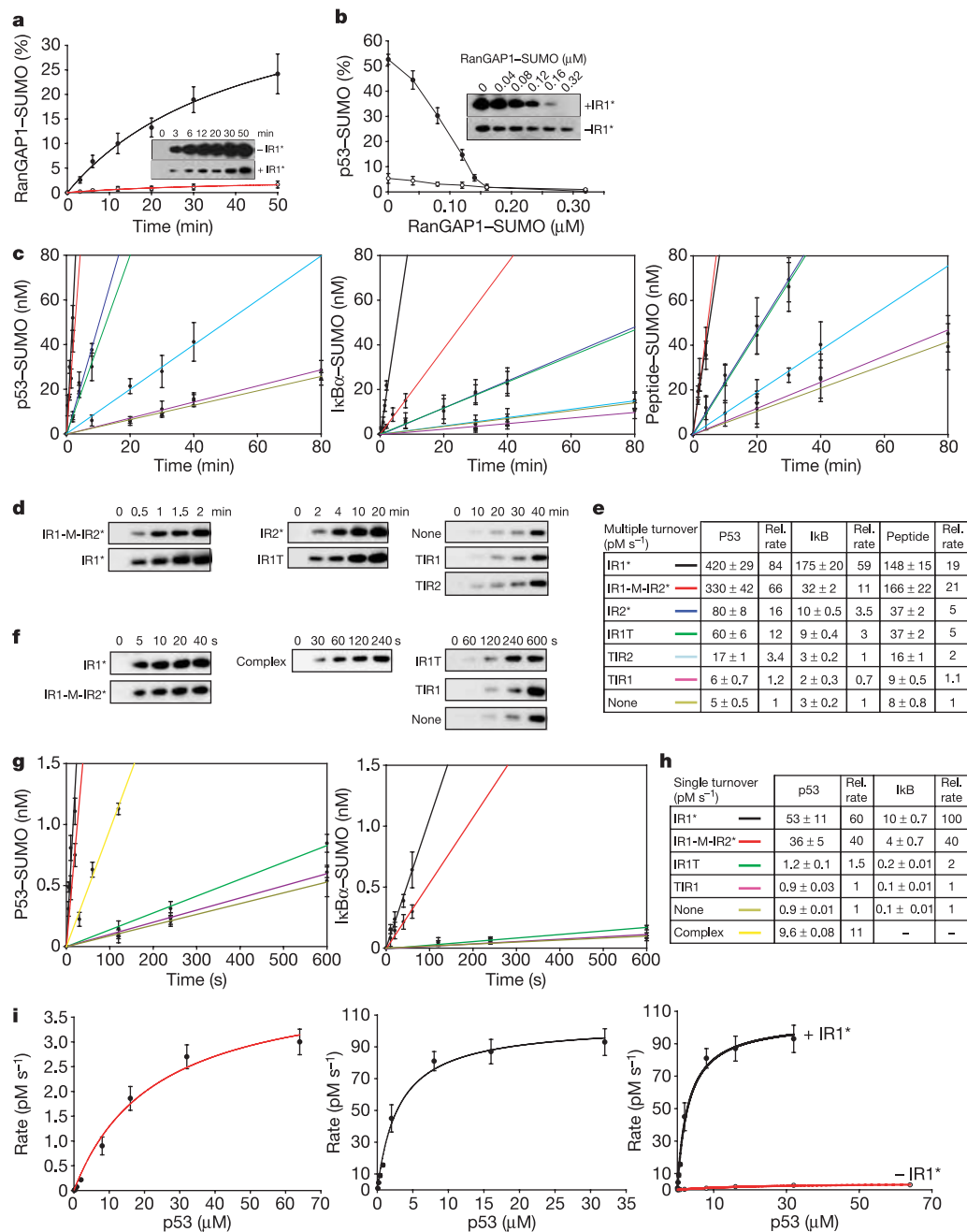


Figure 4 | Nup358/RanBP2 activities. **a**, SUMO-1 RanGAP1 conjugation under multiple turnover with (open circles) or without (filled circles) IR1*; gel insets are shown. **b**, Inhibition of SUMO-1 p53 conjugation by SUMO-1–RanGAP1, with (filled circles) or without (open circles) IR1*; gel insets are shown. **c**, Multiple-turnover SUMO-1 conjugation rates for p53 tetramerization domain (left), IkBα (middle) and p53 peptide (right). **d**, Gel insets for **c** with p53. **e**, Rates (pM s⁻¹) and relative rates for multiple turnover. **f**, Gel insets for **g** using p53. **g**, Single-turnover rates for p53 tetramerization domain (left) and IkBα (right) using Nup358/RanBP2 constructs or IR1-M–RanGAP1–Ubc9–SUMO–Nup358/RanBP2 (yellow). **h**, Rates (pM s⁻¹) for **g**. **i**, Single-turnover rates for p53 without (red) or with (black) IR1*. Note the different scales in the three panels. Results in graphs are means ± s.d.

to that observed in our complex, E3 activity is achieved independently of contacts to the substrate or E2's active site. So how does Nup358/RanBP2 work? Conjugation rates could be increased by an allosteric mechanism that alters the E2 active site indirectly to enhance catalysis¹⁸, possibly activating the thioester, leaving group, or alters E1's ability to charge E2. However, thioester reactivity was not altered when E2-SUMO-thioester was incubated with 0, 1 or 10 mM dithiothreitol (DTT) with or without E3. Differences were also not observed in E1-mediated E2-SUMO-thioester formation with or without E3, although 10:1 or 100:1 E3:E2 molar ratios inhibited the formation of E1-SUMO and E2-SUMO-thioester.

The importance of motif I and motif II for IR1* and IR2* E3 activity and the small interface observed between Ubc9 and SUMO indicated a possible model in which Nup358/RanBP2 catalyses E3 activity by tethering SUMO and Ubc9 together to reduce conformational flexibility, to prevent non-productive E2-SUMO conformations, and to align the complex and thioester for Ubl transfer. If this is true, and contrary to previous observations^{12,18,19} or the lack of substrate contacts, both substrate binding and catalysis should be affected. To test this, initial rate velocities were calculated over various p53 concentrations under single-turnover conditions with and without IR1*. The kinetic data fitted well to an equation from which K_d and k_2 (k_{cat}) were derived (Fig. 4i; see Methods). In the absence of IR1*, K_d and k_2 for p53 conjugation were $23.1 \pm 4.8 \mu\text{M}$ and $4.27 \pm 0.37 \text{ pM s}^{-1}$, respectively ($\pm$ s.d.). In the presence of IR1*, K_d and k_2 were $3.04 \pm 0.47 \mu\text{M}$ and $104.7 \pm 4.3 \text{ pM s}^{-1}$, respectively. Thus, both K_d and k_2 were affected by IR1* to increase K_d/k_2 about 185-fold.

SUMO-conjugated RanGAP1, in stoichiometric quantities, sequesters Ubc9 and substantially diminishes its ability to promote conjugation of exogenous substrate in the presence of Nup358/RanBP2 (see above). Because IR1-M contacted only Ubc9 and SUMO in our complex, we predicted that an additional equivalent of Ubc9-SUMO-thioester should compete with a preformed E2-E3-SUMO-RanGAP1 complex for IR1-M interaction. In fact, single-turnover assays indicated that IR1-M dissociated from the complex, enhancing Ubc9-SUMO-1-thioester conjugation 11-fold (Fig 4f-h). These data indicate that although Nup358/RanBP2 might form a stable complex with SUMO-RanGAP1, it might still be available to promote E3 activity by binding additional E2-SUMO-thioesters, a model that has cellular implications for Nup358/RanBP2 E3 activity if E2-SUMO concentrations are sufficient to compete for Nup358/RanBP2 at the NPC.

The SUMO-RanGAP1-Ubc9-Nup358/RanBP2 structure provides data for several interactions that are crucial for E3-assisted E2 conjugation, including the model that Nup358/RanBP2 enhances conjugation by coordinating the E2-SUMO-thioester optimally for substrate binding and catalysis. Comparison between Nup358/RanBP2-Ubc9-SUMO-RanGAP1 and two other E2-E3 structures^{26,27}, E6AP-UbcH7 and c-Cbl-UbcH7, reveal distinct but overlapping E2 surfaces used in each complex. By aligning respective E2s, SUMO was placed into the other E2-E3 complexes. The RING domain and other E3 surfaces are well positioned to recognize Ub within the modelled E2-Ub-thioester complexes (Supplementary Fig. 3).

The indirect mechanism used by Nup358/RanBP2 to enhance SUMO conjugation might shed some light on 'allosteric' activation observed in other Ub/Ubl pathways such as the RING-finger and Nedd8-induced activation of the Cullin SCF complexes⁷, Apc11 induced activation of the APC complex²⁷, all of which might coordinate E2-Ub in a complex similar to that observed in our structure. Mechanisms proposed here might also provide insights, although less clear ones, into activities that promote E2-Ub/Ubl-thioester activation during polyubiquitin chain formation²⁸.

METHODS

Cloning, expression and protein purification. Preparation of human E1

(Sae1/Sae2), E2 (Ubc9), SUMO-1 (1-97), SUMO-1 (18-97), the C-terminal p53 tetramerization domain and IkB α have been described previously^{9,21}. Human RanGAP1 (residues 419-587) and Nup558/RanBP2 constructs (IR1*, IR1-M, IR1-M-IR2*, TIR1, IR1T, IR2* and TIR2; see Fig. 3a) were amplified by polymerase chain reaction and cloned into a Smt3 vector^{9,21}. Constructs were confirmed by DNA sequencing. E3 constructs included IR1-M-IR2* (residues 2,631-2,771), IR1* (residues 2,631-2,695), IR2* (residues 2,709-2,771), TIR1 (residues 2,640-2,695), TIR2 (residues 2,718-2,771) and IR1T (residues 2,631-2,670). The synthetic p53-derived peptide includes residues 380-393 (HKKLMFKTEGPDSD). Biochemical assays used proteins that were concentrated in buffer containing 350 mM NaCl, 20 mM Tris-HCl pH 8.0 and 1 mM DTT, flash-frozen in liquid nitrogen and stored at -80°C . Cultures were fermented at 37°C to a D_{600} of 3, induced with 0.75 mM isopropyl β -D-thiogalactoside for 4-6 h at 30°C , harvested and suspended in 50 mM Tris-HCl pH 8.0, 20% w/v sucrose, 350 mM NaCl, 20 mM imidazole, 0.1% IGEPAL, 1 mM phenylmethylsulphonyl fluoride, 1 mM 2-mercaptoethanol, $10 \mu\text{g ml}^{-1}$ DNase before sonication and the removal of insoluble material by centrifugation. Proteins were purified by metal-affinity chromatography (Qiagen), gel filtration (Superdex75 or Superdex200; Pharmacia), and ion exchange (MonoQ and MonoS). SUMO-RanGAP1 was prepared and combined with Ubc9 and IR1-M (residues 2,631-2,711), purified by gel filtration (Superdex200), and concentrated to 10 mg ml^{-1} in 50 mM NaCl, 10 mM Tris-HCl pH 8.0, 1 mM DTT, frozen in liquid nitrogen and stored at -80°C .

Crystallographic analysis. Crystals were obtained at 18°C by hanging-drop vapour diffusion against a well solution containing 18% w/v PEG4000, 0.1 M sodium citrate pH 5.0, 0.2 M ammonium acetate and transferred to crystallization solutions containing 12% ethylene glycol before cryoprotection. Data were processed and the structure solved by molecular replacement using human RanGAP1-Ubc9 coordinates to phase the complex at 4 Å. SUMO-1 was modelled on the basis of previous structures²¹. Nup358/RanBP2 was modelled into electron density that became apparent after refinement of SUMO-RanGAP1-Ubc9. Nup358/RanBP2 residues 2,663, 2,665, 2,666, 2,669 and 2,670, not clearly present in the electron density, were modelled without side chains. Refinement and data statistics are provided in the text and Supplementary Table 1.

Biochemical assays. Assays were conducted in triplicate. Samples were removed at specified times, denatured in non-reducing SDS-PAGE buffer containing 4 M urea (single turnover) or reducing SDS-PAGE buffer (multiple turnover), analysed by SDS-PAGE and western blotting with a polyclonal rabbit antibody against SUMO-1 (Boston Biochem), and developed by enhanced chemiluminescence with ECL-Plus (Amersham). Data were imaged with a Fujifilm LAS-3000 Imager and quantified with Image Gauge v4.0 (Fujifilm).

For the multiple-turnover reaction, reactions were performed at 37°C in 150 nM E1, 100 nM Ubc9 (E2), 10 μM SUMO-1 (1-97), 5 mM MgCl_2 , 0.1% Tween 20, 20 mM Hepes pH 7.5, 50 mM NaCl, 1 mM DTT, and the indicated Nup358/RanBP2 constructs (200 nM) using 8 μM p53 tetramerization domain (p53), 4 μM IkB α or 500 μM p53 peptide as substrate. RanGAP1 conjugation was assessed in the presence or absence of IR1* (200 nM). Inhibition of p53 conjugation by RanGAP-SUMO was assessed with and without IR1* (200 nM) in reactions containing RanGAP1-SUMO-1 at 0, 0.04, 0.08, 0.12, 0.14, 0.16 or 0.32 μM .

For the single-turnover reaction, E2-SUMO-thioester was formed at 37°C in non-reducing buffer containing 100 nM E1, 1 μM Ubc9 (E2), 5 mM MgCl_2 , 0.1% Tween 20, 20 mM Hepes pH 7.5, 50 mM NaCl, 1 μM mature SUMO-1 (1-97) and 1 μM ATP, and stopped after 10 min by tenfold dilution at 4°C in buffer containing 5 mM EDTA. Diluted reactions contained 10 nM E1, 100 nM Ubc9 (E2), 100 nM mature SUMO-1, 0.5 mM MgCl_2 , 5 mM EDTA, 0.1% Tween 20, 20 mM Hepes pH 7.5, 50 mM NaCl, either 8 μM p53 or 4 μM IkB α , and the indicated Nup358/RanBP2 elements at 100 nM. Reactions also included p53 in the presence of 100 nM RanGAP1-Ubc9-SUMO-Nup358/RanBP2 (IR1-M). Data in Supplementary Fig. 2 were generated under single-turnover conditions by replacing SUMO-1 with SUMO-2 (1-93) or SUMO-3 (1-92). Rate constants were calculated for p53 by measuring rate velocities under single-turnover conditions with and without IR1* (100 nM), using 0, 0.08, 0.4, 0.8, 2, 4, 16, 32 or 64 μM p53. Data were fitted to a hyperbolic two-parameter single rectangular function to derive reaction constants ($v = k_2[S]/K_d + [S]$, where k_2 is the rate constant, K_d is the dissociation constant, v is velocity and $[S]$ is the substrate concentration). The kinetic scheme used to evaluate K_d (k_{-1}/k_1) and k_2 was $E + S \rightleftharpoons ES \rightarrow E' + P$, where $k_{-1} \gg k_2$, $[S] \gg [E]$, and the chemistry is irreversible²⁹.

Received 11 January; accepted 31 March 2005.

1. Johnson, E. S. Protein modification by SUMO. *Annu. Rev. Biochem.* **73**, 355-382 (2004).

2. Hershko, A. & Ciechanover, A. The ubiquitin system. *Annu. Rev. Biochem.* **67**, 425–479 (1998).
3. Matunis, M. J., Coutavas, E. & Blobel, G. A novel ubiquitin-like modification modulates the partitioning of the Ran-GTPase-activating protein RanGAP1 between the cytosol and the nuclear pore complex. *J. Cell Biol.* **135**, 1457–1470 (1996).
4. Mahajan, R., Delphin, C., Guan, T., Gerace, L. & Melchior, F. A small ubiquitin-related polypeptide involved in targeting RanGAP1 to nuclear pore complex protein RanBP2. *Cell* **88**, 97–107 (1997).
5. Saitoh, H., Pu, R., Cavenagh, M. & Dasso, M. RanBP2 associates with Ubc9p and a modified form of RanGAP1. *Proc. Natl Acad. Sci. USA* **94**, 3736–3741 (1997).
6. Zhang, H., Saitoh, H. & Matunis, M. J. Enzymes of the SUMO modification pathway localize to filaments of the nuclear pore complex. *Mol. Cell. Biol.* **22**, 6498–6508 (2002).
7. Deshaies, R. J. SCF and Cullin/Ring H2-based ubiquitin ligases. *Annu. Rev. Cell Dev. Biol.* **15**, 435–467 (1999).
8. Huibregtse, J. M., Scheffner, M., Beaudenon, S. & Howley, P. M. A family of proteins structurally and functionally related to the E6-AP ubiquitin-protein ligase. *Proc. Natl Acad. Sci. USA* **92**, 2563–2567 (1995).
9. Bernier-Villamor, V., Sampson, D. A., Matunis, M. J. & Lima, C. D. Structural basis for E2-mediated SUMO conjugation revealed by a complex between ubiquitin-conjugating enzyme Ubc9 and RanGAP1. *Cell* **108**, 345–356 (2002).
10. Johnson, E. S. & Gupta, A. A. An E3-like factor that promotes SUMO conjugation to the yeast septins. *Cell* **106**, 735–744 (2001).
11. Kahyo, T., Nishida, T. & Yasuda, H. Involvement of PIAS1 in the sumoylation of tumor suppressor p53. *Mol. Cell* **8**, 713–718 (2001).
12. Pichler, A., Gast, A., Seeler, J. S., Dejean, A. & Melchior, F. The nucleoporin RanBP2 has SUMO1 E3 ligase activity. *Cell* **108**, 109–120 (2002).
13. Kagey, M. H., Melhuish, T. A. & Wotton, D. The polycomb protein Pc2 is a SUMO E3. *Cell* **113**, 127–137 (2003).
14. Yokoyama, N. *et al.* A giant nucleopore protein that binds Ran/TC4. *Nature* **376**, 184–188 (1995).
15. Wu, J., Matunis, M. J., Kraemer, D., Blobel, G. & Coutavas, E. Nup358, a cytoplasmically exposed nucleoporin with peptide repeats, Ran-GTP binding sites, zinc fingers, a cyclophilin A homologous domain, and a leucine-rich region. *J. Biol. Chem.* **270**, 14209–14213 (1995).
16. Joseph, J., Liu, S. T., Jablonski, S. A., Yen, T. J. & Dasso, M. The RanGAP1-RanBP2 complex is essential for microtubule-kinetochore interactions *in vivo*. *Curr. Biol.* **14**, 611–617 (2004).
17. Saitoh, H., Pizzi, M. D. & Wang, J. Perturbation of SUMOylation enzyme Ubc9 by distinct domain within nucleoporin RanBP2/Nup358. *J. Biol. Chem.* **277**, 4755–4763 (2002).
18. Pichler, A., Knipscheer, P., Saitoh, H., Sixma, T. K. & Melchior, F. The RanBP2 SUMO E3 ligase is neither HECT nor RING type. *Nature Struct. Mol. Biol.* **11**, 984–991 (2004).
19. Tatham, M. H. *et al.* Unique binding interactions among Ubc9, SUMO and RanBP2 reveal a mechanism for SUMO paralog selection. *Nature Struct. Mol. Biol.* **12**, 67–74 (2004).
20. Nicholls, A., Sharp, K. A. & Honig, B. Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons. *Proteins* **11**, 281–296 (1991).
21. Lois, L. M. & Lima, C. D. Structures of the Small Ubiquitin-like MODifier E1 activating enzyme provide insights into SUMO activation and the mechanism for E2 recruitment to E1. *EMBO J.* **24**, 439–451 (2005).
22. Hamilton, K. S. *et al.* Structure of a conjugating enzyme-ubiquitin thiolester intermediate reveals a novel role for the ubiquitin tail. *Structure* **9**, 897–904 (2001).
23. Wu, P. Y. *et al.* A conserved catalytic residue in the ubiquitin-conjugating enzyme family. *EMBO J.* **22**, 5241–5250 (2003).
24. Song, J., Durrin, L. K., Wilkinson, T. A., Krontiris, T. G. & Chen, Y. Identification of a SUMO-binding motif that recognizes SUMO-modified proteins. *Proc. Natl Acad. Sci. USA* **101**, 14373–14378 (2004).
25. Hannich, J. T. *et al.* Defining the SUMO-modified proteome by multiple approaches in *Saccharomyces cerevisiae*. *J. Biol. Chem.* **280**, 4102–4110 (2005).
26. Huang, L. *et al.* Structure of an E6AP-UbcH7 complex: insights into ubiquitination by the E2–E3 enzyme cascade. *Science* **286**, 1321–1326 (1999).
27. Leversen, J. D. *et al.* The APC11 RING-H2 finger mediates E2-dependent ubiquitination. *Mol. Biol. Cell* **11**, 2315–2325 (2000).
28. Pickart, C. M. Mechanisms underlying ubiquitination. *Annu. Rev. Biochem.* **70**, 503–533 (2001).
29. Strickland, S., Palmer, G. & Massey, V. Determination of dissociation constants and specific rate constants of enzyme–substrate (or protein–ligand) interactions from rapid reaction kinetic data. *J. Biol. Chem.* **250**, 4048–4052 (1975).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank M. J. Matunis for the original clone containing Nup358/RanBP2 (residues 2,596–2,836), and K. R. Rajashankar and A. Yunus for discussion and for reagents that contributed to this work. Use of the Advanced Photon Source (APS) is supported by the US Department of Energy, Office of Science, Office of Basic Energy Sciences. Use of the SGX Collaborative Access Team beamline facilities at Sector 31 of the APS was provided by Structural Genomics, Inc., which constructed and operates the facility. D.R. and C.D.L. were supported in part by a National Institutes of Health grant. C.D.L. acknowledges support from the Rita Allen Foundation.

Author Information Coordinates have been deposited with the Protein Data Bank under accession number 1Z5S. Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to C.D.L. (limac@mskcc.org).

Structural basis for nuclear import complex dissociation by RanGTP

Soo Jae Lee, Yoshiyuki Matsuura, Sai Man Liu & Murray Stewart

Nuclear protein import is mediated mainly by the transport factor importin- β that binds cytoplasmic cargo, most often via the importin- α adaptor, and then transports it through nuclear pore complexes. This active transport is driven by disassembly of the import complex by nuclear RanGTP^{1–4}. The switch I and II loops of Ran change conformation with nucleotide state^{5–7}, and regulate its interactions with nuclear trafficking components. Importin- β consists of 19 HEAT repeats that are based on a pair of antiparallel α -helices (referred to as the A- and B-helices). The HEAT repeats stack to yield two C-shaped arches, linked together to form a helicoidal molecule that has considerable conformational flexibility^{1,8–12}. Here we present the structure of full-length yeast importin- β (Kap95p or karyopherin- β) complexed with RanGTP, which provides a basis for understanding the crucial cargo-release step of nuclear import. We identify a key interaction site where the RanGTP switch I loop binds to the carboxy-terminal arch of Kap95p. This interaction produces a change in helicoidal pitch that locks Kap95p in a conformation that cannot bind importin- α or cargo. We suggest an allosteric mechanism for nuclear import complex disassembly by RanGTP.

RanGTP ultimately provides the driving force for nuclear protein import⁴, and the way in which the interaction between Ran and importin- β is modulated by the state of the bound nucleotide (GTP or GDP) provides the basis for two key steps in nuclear protein import. In the nucleus, high-affinity RanGTP binding dissociates the importin-cargo complex, making import irreversible and recycling the importins to the cytoplasm. RanGAP-mediated GTP hydrolysis in the cytoplasm generates RanGDP, which dissociates from importin- β , freeing it to undertake another import cycle^{1–4}. The crystal structure of a complex between Ran-GppNHp (GppNHp is a non-hydrolysable GTP analogue) and a fragment of importin- β containing the amino-terminal arch (HEAT repeats 1–11) (ref. 6) showed that the switch II loop of Ran binds to an N-terminal CRIME motif in HEAT repeats 1–4, and that a basic patch on Ran (residues Arg 140, Lys 141, Lys 159 and Arg 166) binds to an acidic loop in HEAT repeat 8 of importin- β . The conformational change associated with nucleotide state is much greater in the switch I loop than the switch II loop^{5–7}, and so it was paradoxical that the switch I loop did not appear to bind importin- β directly. However, the switch I loop in RanGDP would clash with HEAT repeat 1, at least partially accounting for the markedly lower affinity of RanGDP⁶.

The 2.7 Å resolution crystal structure of full-length Kap95p complexed with residues 1–176 of RanGTP (Fig. 1a) showed a new Ran-binding site in the C-terminal arch of Kap95p, in addition to the sites already identified in the N-terminal arch⁶. This additional site in Kap95p interacts intimately with the switch I loop of RanGTP, with Arg 29^{Ran} and Lys 37^{Ran} forming salt bridges and hydrogen bonds with HEAT repeats 12–15, and Phe 35^{Ran} participating in hydrophobic interactions (Fig. 1b, c), while Lys 152^{Ran}, Asn 154^{Ran}, Asn 156^{Ran} and Phe 157^{Ran} also contribute to the interface. A key

feature of this interaction is that Lys 37^{Ran} and Lys 152^{Ran} form salt bridges with a cluster of acidic residues in HEAT repeat 14 (Glu 615, Asp 616 and Asp 617) of Kap95p. In addition, Arg 29^{Ran} and Asn 154^{Ran} form an extensive network of hydrogen bonds with HEAT repeat 13 (Fig. 1c). In contrast with the two sites in the N-terminal arch, where RanGTP binds to the B-helices of the inner concave surface of Kap95p (see Supplementary Fig. 1), the site in the C-terminal arch involves primarily one edge of the helicoid and residues in the linkers between A- and B- helices within HEAT repeats.

Overall, this third site buried 1,530 Å² of surface area, compared with 2,786 Å² for the two sites in the N-terminal arch, and provides a rationale for why RanGTP binds more strongly than RanGDP. The energetic difference between the RanGTP and RanGDP states is estimated⁶ to be of the order of 22 kJ mol^{–1}. The structure of the Kap95p–RanGTP complex shows that, in addition to the steric clash with RanGDP noted previously⁶, the direct interaction between the Ran switch I loop and Kap95p makes a major positive contribution to the affinity of Kap95p for RanGTP. The energy from RanGTP binding to the C-terminal arch might be required to compensate for distortions introduced into Kap95p by complex formation, or to compensate for the loss of entropy resulting from its reduced flexibility. This might account for the observation¹³ that, whereas the N-terminal arch (residues 1–462) of human importin- β binds RanGTP with an affinity comparable to the full-length molecule, fragments that contain the N-terminal arch and only part of the C-terminal arch (for example, residues 1–618) show lower affinity.

We showed that the C-terminal binding site of Kap95p is important for import complex formation by using RanGTP containing K37D and K152A mutations, which disrupt the formation of the key salt bridges between RanGTP and the acidic cluster in HEAT repeat 14 (Fig. 1c). RanGTP, but not RanGDP, disrupts the interaction between Kap95p and the IBB (importin- β -binding) domain of yeast importin- α (Kap60p) by forming a Kap95p–RanGTP complex (Fig. 2). Previous work¹⁴ has demonstrated that K37D/K152A–RanGTP retains the ability to bind yeast Ran binding protein 1 (RanBP1 or Yrb1p), showing that the mutant retains the ability to bind to other proteins in the same way as wild-type RanGTP. Moreover, because this mutant retains the residues involved in high-affinity binding to the Kap95p N-terminal arch, it is still able to bind Kap95p (Fig. 2a). However, whereas both wild-type and Q69L–RanGTP completely released Kap95p from the Kap60p IBB domain, K37D/K152A–Ran was markedly less efficient and most of the Kap95p remained bound to the IBB domain (Fig. 2b). Thus, the interaction between the Ran switch I loop and the C-terminal arch of Kap95p is required for effective dissociation of the Kap60p–Kap95p import complex. We obtained similar results using human importin- β , indicating that it interacts with RanGTP in the same manner as Kap95p (Fig. 2a, b).

RanGTP binding produced a substantial conformational change

in full-length Kap95p (Fig. 1d). In the complex, Kap95p is constructed from the same helicoidal stacking of HEAT repeats seen in mammalian importin- β ^{8,10}, but both the shape and the relative positions of the N- and C-terminal arches are changed. This alters the local helicoidal pitch of Kap95p in complex with RanGTP, compared with its homologue importin- β bound to import sub-

strates such as the IBB domain or the sterol regulatory element binding protein (SREBP-2). These changes are most pronounced at the termini of Kap95p, whereas the central region of the molecule, comprising HEAT repeats 4–11, is remarkably similar to importin- β bound to substrates. For example, the r.m.s. C α deviation of structurally equivalent residues over HEAT repeats 4–11 is only

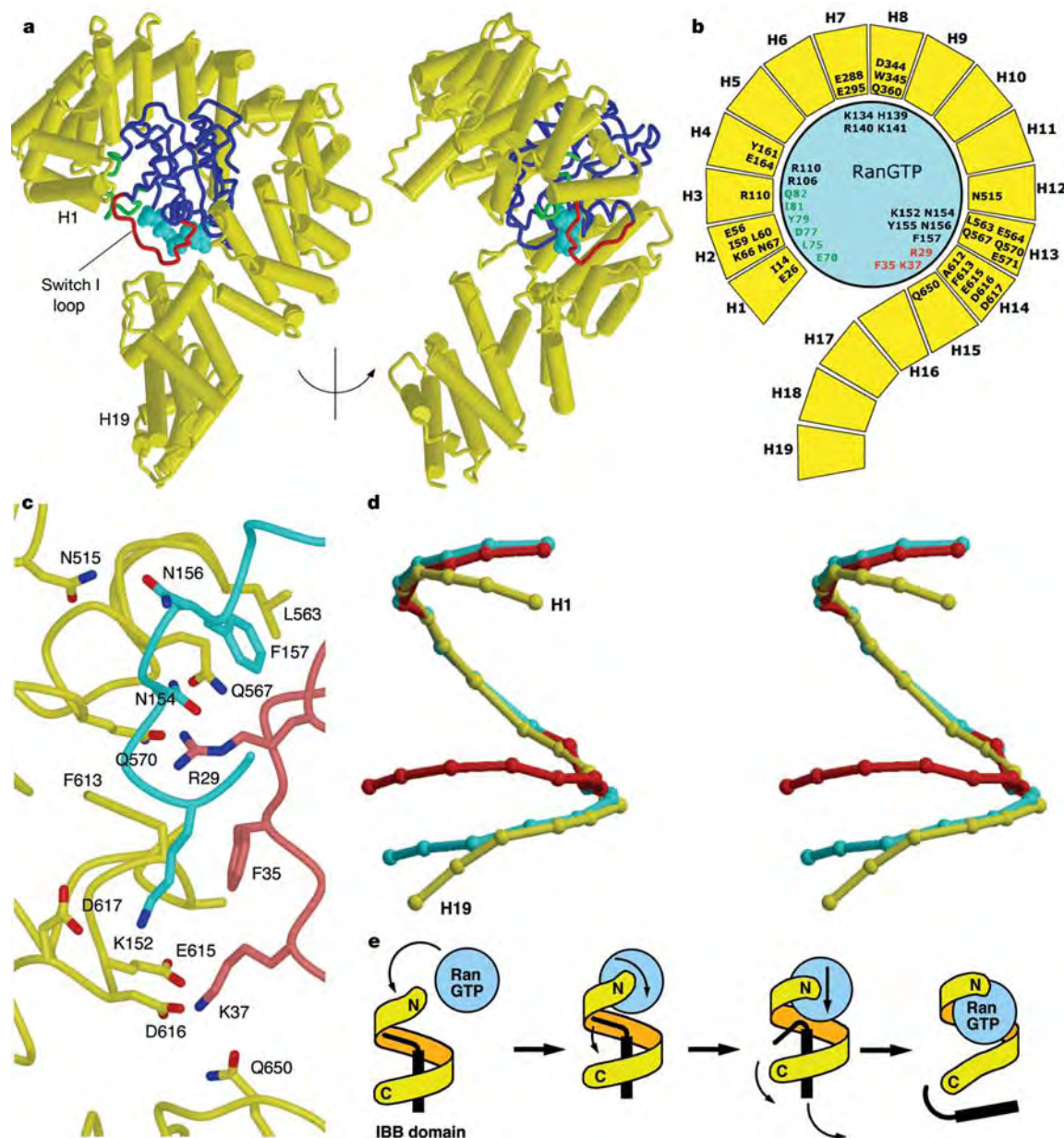


Figure 1 | Structure of the Kap95p-RanGTP complex. **a**, Two views of the Kap95p-RanGTP complex, rotated by 90° relative to one another. Kap95p is shown in yellow and the main-chain trace of RanGTP is shown in dark blue, with the switch I loop in red and the switch II loop in green. The bound GTP is shown in space-filling format. **b**, Schematic illustration of the residues interacting at each site. The CRIME motif in HEAT repeats (H)1–4 of Kap95p interacts primarily with the switch II loop of RanGTP, and the basic patch on RanGTP binds mainly to the acidic loop in HEAT repeat 8. The third interaction site primarily involves interactions between the switch I loop (red) and residues in HEAT repeats 12–15. **c**, Details of the interaction site in the Kap95p C-terminal arch (yellow) that involves the Ran switch I loop (red) and Ran residues 151–158 (blue). Lys 37 and Lys 152 of Ran make salt bridges with acidic residues from HEAT repeat 14; Arg 29^{Ran}, Asn 154^{Ran} and Asn 156^{Ran} form an extensive hydrogen-bonded network; and Phe 35^{Ran} and Phe 157^{Ran}, together with Phe 613 and Leu 563 of Kap95p, are also

buried. **d**, Stereo view showing the conformational changes between Kap95p bound to different partners, IBB (red), SREBP-2 (cyan) and RanGTP (yellow), indicative of considerable molecular flexibility. The centres of consecutive HEAT repeats in each Kap95p molecule are represented by spheres. RanGTP binding locks Kap95p into a conformation with a higher helicoidal pitch that is incompatible with binding nuclear import partners. **e**, Schematic illustration of an unzipping mechanism by which RanGTP might displace the IBB domain from Kap95p. Kap95p is shown in yellow, and RanGTP in blue. An extended section of the L-shaped IBB domain (black) binds Kap95p HEAT repeats 7–11, while an α -helical segment binds HEAT repeats 12–19. RanGTP binds initially to the CRIME domain at the Kap95p N terminus and then displaces the extended IBB region by binding to the acidic loop in HEAT repeat 8. RanGTP binding to the additional site in the C-terminal arch then increases the helicoidal pitch of Kap95p, resulting in a mismatch between it and the IBB domain helix, which is then released.

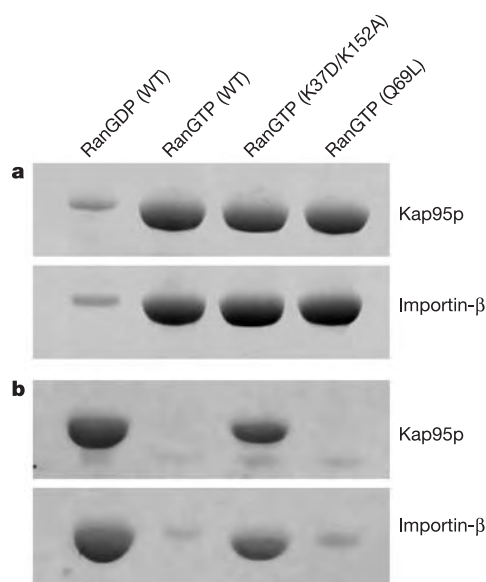


Figure 2 | RanGTP with K37D/K152A mutations binds Kap95p and importin-β but does not displace the IBB domain. **a**, Although both Kap95p and human importin-β only bind weakly to wild-type (WT) RanGDP, they bind K37D/K152A-RanGTP, Q69L-RanGTP and WT RanGTP with comparable affinity. **b**, The amount of Kap95p and human importin-β bound to GST-IBB depends on the state of the nucleotide bound to Ran. Kap95p and importin-β remain bound in the presence of RanGDP, but are displaced by RanGTP and Q69L-RanGTP. However, Kap95p and importin-β are displaced much less efficiently by the GTP-bound form of the Ran K37D/K152A mutant, which disrupts the interaction at the additional Ran-binding site in the C-terminal arch of Kap95p (Fig. 1c).

1.6 Å between the importin-β-IBB complex⁸ (PDB accession number 1QGK) and the Kap95p-RanGTP complex. However, in the Kap95p-RanGTP complex, the curvature of HEAT repeats 1–3 is increased and that of HEAT repeats 12–19 is decreased (Fig. 1d) relative to the importin-β-IBB complex. In both areas, this change appears to be accommodated mainly by relatively small changes in orientation between adjacent HEAT repeats. Although their internal structure remains essentially unchanged, cumulative displacements result in the C terminus of Kap95p bound to RanGTP being displaced by over 20 Å relative to the importin-β-IBB complex.

Although other karyopherin/importin-β family members appear to bind the Ran switch II loop through the conserved N-terminal domain (HEAT repeats 1–4), there is considerable variation in the additional RanGTP-binding sites on these molecules. For example, in the importin-β family members Kap-β2 (ref. 5) and

CRM1/Xpo1p (ref. 15), RanGTP binding appears to also involve a large loop near the centre of the molecule. However, in its complex with Kap60p and RanGTP, the exportin Cse1p binds to a region near the centre of the molecule as well as to a long loop in HEAT repeat 19 that extends back along the inner concave surface of the helixoid¹⁴.

Crystal structures of importin-β bound to three different cargoes have shown marked differences in the structural determinants by which cargoes are recognized, as well as changes in helicoidal pitch of importin-β, indicative of molecular flexibility^{8,10,16}. The IBB domain of importin-α has a particularly extensive interaction interface that involves importin-β HEAT repeats 7–19 and is dominated by electrostatic interactions between basic IBB domain residues and acidic residues in the B-helices that form the inner concave surface of the molecule and form an almost continuous acidic patch⁸. The IBB domain forms an L-shaped structure, with an α-helix interacting with HEAT repeats 12–19, and an extended section at right-angles that binds to HEAT repeats 7–11 as well as the acidic loop in HEAT repeat 8. In contrast, SREBP-2 (ref. 10) binds primarily to the long α-helices of HEAT repeats 7 and 17. Parathyroid hormone related protein (PTHrP) has an extended binding site on importin-β that spans HEAT repeats 2–11 in the N-terminal arch and largely overlaps the RanGTP binding domain¹⁶. Low-angle X-ray studies¹¹ are also consistent with karyopherins having considerable flexibility and altering conformation upon binding RanGTP or cargo. Thus, by using different binding sites and different conformations, importin-β/Kap95p is able to bind a broad range of different cargoes. How then is RanGTP binding able to release all these different cargoes?

The structure of the Kap95p-RanGTP complex suggests an allosteric mechanism by which RanGTP binding generates a conformational change in the molecule that alters the binding sites for cargoes. Although the flexibility of Kap95p/importin-β enables it to bind different cargoes essentially by an induced-fit mechanism, we propose that, by binding to sites in both the C- and N-terminal arches, RanGTP locks Kap95p into a conformation that is unable to bind cargo. Because the binding constant of cargoes is generally in the range of 10 nM (ref. 4), it is likely that RanGTP actively displaces the cargo, rather than passively binding free Kap95p formed by reversible import complex dissociation. Because the binding site for the IBB domain is so extensive, RanGTP might displace it in a series of steps—essentially unzipping the Kap95p-IBB domain interface. Figure 1e illustrates one way in which this could occur. RanGTP might first bind to the conserved CRIME domain at the Kap95p N terminus (HEAT repeats 1–4), which does not interact directly with the IBB domain. Next, RanGTP could displace the extended region of the IBB domain from HEAT repeats 7–11, especially through the interaction between the Ran basic patch and the negatively charged loop of HEAT repeat 8. The binding site for the RanGTP switch I loop is on the side of Kap95p, and so does not overlap with residues involved in IBB binding, which are located primarily along the inner concave surface of the molecule⁸. However, RanGTP binding to the site in the C-terminal arch would induce a conformational change that increases the local helicoidal pitch, breaking the structural complementarity between Kap95p and the α-helical region of the IBB domain, so that they can no longer interact effectively. A steric clash with RanGTP is probably not important for displacing SREBP-2, but might be more important for PTHrP¹⁶. SREBP-2 is gripped like chopsticks by the long helices of HEAT repeats 7 and 17 (ref. 10). In the Kap95p-RanGTP structure, the tips of these helices move apart (see Supplementary Fig. 2) and so would be unable to bind the cargo. Other karyopherin/importin-β family members might use analogous, RanGTP-based allosteric mechanisms to regulate cargo binding. Low-angle X-ray scattering indicates a considerable variation between different karyopherins bound to RanGTP and/or different substrates¹¹, suggesting that the detailed way in which RanGTP binding generates these allosteric changes might differ between family members.

In summary, the structure of the Kap95p-RanGTP complex

Table 1 | Crystallographic statistics

Data collection	Space group	<i>P</i> 2 ₁ 2 ₁
	Unit cell dimensions (Å)	<i>a</i> = 107.0; <i>b</i> = 127.9; <i>c</i> = 161.7
	Resolution range (Å)*	20.0–2.70 (2.85–2.70)
	Total/unique observations*	237,350 (34,147)/61,240 (8,829)
	Completeness (%)	99.5 (99.7)
	<i>R</i> _{pim} (%)	6.7 (37.7)
	Mean/ <i>σ</i> (I)*	10.0 (2.0)
Refinement	<i>R</i> _{cryst} / <i>R</i> _{free} (%)	22.3/27.9
	Bond length r.m.s.d. (Å)	0.009
	Bond angle r.m.s.d. (°)	1.17
	Ramachandran plot (%)	
	Core region	91.8
	Allowed	7.6
	Generously allowed	0.6
	Forbidden	0

*Parentheses refer to final resolution shell.

provides a basis for understanding how RanGTP dissociates the major nuclear protein import complex, by locking Kap95p in a conformation in which it is unable to exploit its flexibility to bind to different partners.

METHODS

Preparation of the Kap95p–RanGTP complex. Untagged Kap95p and glutathione *S*-transferase (GST)–RanGTP (residues 1–176) were expressed separately in *Escherichia coli* strain BL21-RIL(DE3) (Stratagene) at 20 °C, as described previously^{14,17}. After collection, the two sets of cells were mixed, suspended in buffer A (50 mM Tris–HCl pH 8.0, 500 mM NaCl, 2 mM MgCl₂, 1 mM dithiothreitol) and lysed by high pressure homogenization at 4 °C. All subsequent steps were performed at 4 °C. Clarified lysates were loaded on glutathione Sepharose 4B resin (Pharmacia) for 3 h and washed with buffer A. After cleavage of the GST tag with thrombin (Sigma), the complex was further purified over Superdex 200 (Pharmacia).

Crystallization. Crystals of the Kap95p–RanGTP complex were grown by vapour diffusion. Proteins were concentrated by centrifugation to a final concentration of 9 mg ml^{−1} in 20 mM Tris–HCl pH 8.0, 150 mM NaCl, 2 mM MgCl₂ and 1 mM dithiothreitol. The reservoir contained 100 mM MES buffer (pH 6.2) with 14–16% PEG 3350 and 1–2 mM MnCl₂. Rod-shaped crystals appeared in two days, and after five days at 20 °C reached maximum dimensions of 70 × 70 × 300 μm.

Data collection and analysis. Crystals were flash-cooled in a 100 K nitrogen gas stream immediately after cryoprotection by addition of 25% (w/v) glycerol to the reservoir solution. Data was collected at beam line 10.1 at the synchrotron radiation source (SRS) in Daresbury, UK. The crystal belonged to orthorhombic space group *P*₂₁₂₁ with unit cell dimensions of *a* = 107.0, *b* = 127.9, *c* = 161.7 Å. The asymmetric unit contained two Kap95p–RanGTP complexes. The data were processed and integrated by MOSFLM and the CCP4 suite¹⁸. Data collection statistics are summarized in Table 1.

Structure solution. We obtained a 2.7-Å resolution structure (Table 1) by molecular replacement¹⁸ using the structure of importin-β residues 1–445 complexed with Ran–GppNHp⁶ to give initial phases. The importin-β sequence in this fragment was replaced with Kap95p and then the remainder of the Kap95p chain was built into solvent-flipped maps one HEAT repeat at a time. After rigid body refinement of individual HEAT repeats using CNS¹⁹, *R*_{free} was 31%. Local rebuilding, especially at the junctions between HEAT repeats, and refinement with Refmac5 (ref. 18) produced a structural model with a crystallographic *R*-factor of 22.3% (*R*_{free} = 27.9%) and good stereochemistry (Table 1). Structural figures were produced using Molscript²⁰ and Raster3D²¹.

Binding assays. Binding of wild-type and mutant RanGTP to Kap95p was essentially as described¹⁴. For the IBB displacement assays, wild-type, Q69L- and K37D/K152A-Ran, human importin-β and GST-IBB were prepared as described^{14,22,23} and charged with GTP either as described¹⁴ or by mixing with His-tagged RCC1 in binding buffer (PBS containing 5 mM MgCl₂, 0.1% Tween20, 2 mM dithiothreitol) containing 10 mM GTP. Both methods gave equivalent results. GST-IBB was immobilized on glutathione Sepharose, washed, and the beads were then incubated with Kap95p to produce the GST-IBB–Kap95p complex. The beads were then incubated with RanGTP (wild type, Q69L or K37D/K152A mutant) in a total volume of 1 ml for 1 h at 4 °C, centrifuged and washed with 2 × 1 ml binding buffer. Bound proteins were eluted with SDS sample buffer and analysed by SDS–polyacrylamide gel electrophoresis (PAGE).

Received 17 February; accepted 1 April 2005.

Published online 1 May 2005.

1. Chook, Y. M. & Blobel, G. Karyopherins and nuclear import. *Curr. Opin. Struct. Biol.* **11**, 703–715 (2001).
2. Mosammaparast, N. & Pemberton, L. F. Karyopherins: from nuclear-transport mediators to nuclear-function regulators. *Trends Cell Biol.* **14**, 547–556 (2004).
3. Weis, K. Regulating access to the genome: nucleocytoplasmic transport throughout the cell cycle. *Cell* **112**, 441–451 (2003).

4. Görlich, D., Seewald, M. J. & Ribbeck, K. Characterization of Ran-driven cargo transport and the RanGTPase system by kinetic measurements and computer simulation. *EMBO J.* **22**, 1088–1100 (2003).
5. Chook, Y. M. & Blobel, G. Structure of the nuclear transport complex karyopherin-β2–Ran–GppNHp. *Nature* **399**, 230–237 (1999).
6. Vetter, I. R., Arndt, A., Kutay, U., Görlich, D. & Wittinghofer, A. Structural view of the Ran–Importin β interaction at 2.3 Å resolution. *Cell* **97**, 635–646 (1999).
7. Vetter, I. R., Nowak, C., Nishimoto, T., Kuhlmann, J. & Wittinghofer, A. Structure of a Ran-binding domain complexed with Ran bound to a GTP analogue: implications for nuclear transport. *Nature* **398**, 39–46 (1999).
8. Cingolani, G., Petosa, C., Weis, K. & Müller, C. W. Structure of importin-β bound to the IBB domain of importin-α. *Nature* **399**, 221–229 (1999).
9. Conti, E. & Izaurralde, E. Nucleocytoplasmic transport enters the atomic age. *Curr. Opin. Cell Biol.* **13**, 310–319 (2001).
10. Lee, S. J. *et al.* The structure of importin-β bound to SREBP-2: nuclear import of a transcription factor. *Science* **302**, 1571–1575 (2003).
11. Fukuhara, N., Fernandez, E., Ebert, J., Conti, E. & Svergun, D. Conformational variability of nucleocytoplasmic transport factors. *J. Biol. Chem.* **279**, 2176–2181 (2004).
12. Cingolani, G., Lashuel, H. A., Gerace, L. & Müller, C. W. Nuclear import factors importin α and importin β undergo mutually induced conformational changes upon association. *FEBS Lett.* **484**, 291–298 (2000).
13. Kutay, U., Izaurralde, E., Bischoff, F. R., Mattaj, J. & Görlich, D. Dominant-negative mutants of importin-β block multiple pathways of import and export through the nuclear pore complex. *EMBO J.* **16**, 1153–1163 (1997).
14. Matsuura, Y. & Stewart, M. Structural basis for the assembly of a nuclear export complex. *Nature* **432**, 872–877 (2004).
15. Petosa, C. *et al.* Architecture of CRM1/Exportin1 suggests how cooperativity is achieved during formation of a nuclear export complex. *Mol. Cell* **16**, 761–775 (2004).
16. Cingolani, G., Bednenko, J., Gillespie, M. T. & Gerace, L. Basis for the recognition of a nonclassical nuclear localization signal by importin β. *Mol. Cell* **10**, 1345–1353 (2002).
17. Bayliss, R., Littlewood, T., Strawn, L. A., Wente, S. R. & Stewart, M. GLFG and FxFG nucleoporins bind to overlapping sites on importin-β. *J. Biol. Chem.* **277**, 50597–50606 (2002).
18. Collaborative Computational Project Number 4, The CCP4 suite: programs for protein crystallography. *Acta Crystallogr. D* **50**, 760–763 (1994).
19. Brünger, A. T. *et al.* Crystallography and NMR system: a new software suite for macromolecular structure determination. *Acta Crystallogr. D* **50**, 905–921 (1998).
20. Kraulis, P. J. MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallogr.* **24**, 946–950 (1991).
21. Merritt, E. A. & Bacon, D. J. Raster3D: photorealistic molecular graphics. *Methods Enzymol.* **277**, 505–524 (1997).
22. Matsuura, Y., Lange, A., Harriman, M., Corbett, A. H. & Stewart, M. Structural basis for the function of Nup2p in cargo release and karyopherin recycling in nuclear import. *EMBO J.* **22**, 5358–5369 (2003).
23. Stewart, M., Kent, H. M. & McCoy, A. J. The structure of the Q69L mutant of GDP-Ran shows a major conformational change in the switch II loop that accounts for its failure to bind nuclear transport factor 2 (NTF2). *J. Mol. Biol.* **284**, 1517–1527 (1998).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank M. Allen and M. Canci for data collection, Y. Yoneda and T. Sekimoto for plasmids, and our colleagues in Cambridge, especially N. Marshall, R. Grant, R. A. Crowther, R. Henderson, K. Nagai and N. Unwin, for their assistance, comments and suggestions. S.J.L. holds an EMBO Long-Term Fellowship and S.M.L. holds an MRC studentship. This work was supported in part by a Human Frontier Research Program grant.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to M.S. (ms@mrc-lmb.cam.ac.uk). Atomic coordinates and structure factors have been deposited in the Protein Data Bank under accession codes 2BKU and r2bkusf, respectively.

Structure of oxidized α -haemoglobin bound to AHSP reveals a protective mechanism for haem

Liang Feng¹, Suiping Zhou², Lichuan Gu¹, David A. Gell³, Joel P. Mackay³, Mitchell J. Weiss², Andrew J. Gow² & Yigong Shi¹

The synthesis of haemoglobin A (HbA) is exquisitely coordinated during erythrocyte development to prevent damaging effects from individual α - and β -subunits^{1,2}. The α -haemoglobin-stabilizing protein (AHSP) binds α -haemoglobin (α Hb), inhibits the ability of α Hb to generate reactive oxygen species and prevents its precipitation on exposure to oxidant stress^{3–5}. The structure of AHSP bound to ferrous α Hb is thought to represent a transitional complex through which α Hb is converted to a non-reactive, hexacoordinate ferric form⁵. Here we report the crystal structure of this ferric α Hb–AHSP complex at 2.4 Å resolution. Our findings reveal a striking bis-histidyl configuration in which both the proximal and the distal histidines coordinate the haem iron atom. To attain this unusual conformation, segments of α Hb undergo drastic structural rearrangements, including the repositioning of several α -helices. Moreover, conversion to the ferric bis-histidine configuration strongly and specifically inhibits redox chemistry catalysis and haem loss from α Hb. The observed structural changes, which impair the chemical reactivity of haem iron, explain how AHSP stabilizes α Hb and prevents its damaging effects in cells.

Free α Hb is toxic because its haem iron can participate in redox chemistry that generates reactive oxygen species (ROS)⁶. Moreover, α Hb is structurally unstable and tends to denature, releasing toxic α -globin polypeptide, and also free haem and iron, which themselves catalyse ROS production. These effects contribute to the pathophysiology of β -thalassaemia, a common inherited anaemia in which mutations that inhibit β -globin synthesis cause the accumulation of free α Hb^{7,8}. AHSP-null red blood cells contain excessive ROS with signs of oxidative damage, haemoglobin precipitates and decreased circulating lifespan^{3,4}. Recently, we showed that the binding of AHSP to oxygenated ferrous (FeII) α Hb induces a unique rearrangement in which the FeII–haem group becomes coordinated by the distal (His 58) but not the proximal (His 87) histidine, which normally binds haem iron in HbA (ref. 5). We proposed that this complex represents a transitional structure because the FeII–haem group was rapidly converted to an oxidized (ferric; FeIII) state in which all six coordinate positions were ligand-bound⁵.

The FeIII– α Hb–AHSP complex was completely oxidized by potassium ferricyanide. To facilitate crystallization, we introduced a point mutation (P30A) into AHSP and removed the carboxy-terminal 11 residues of AHSP, which are known to be dispensable for α Hb binding^{9,10}. We crystallized the FeIII– α Hb–AHSP complex and determined its structure at 2.4 Å resolution (Fig. 1, Table 1 and Supplementary Table 1). There are two α Hb–AHSP complexes in each asymmetric unit. Because these two complexes have identical structural features, we limit our discussion to one such complex.

The FeIII– α Hb–AHSP complex is exclusively in an α -helical conformation (Fig. 1). In the crystals AHSP adopts an elongated

three-helix bundle, whereas α Hb is composed of seven α -helices, known as helices A–C and E–H¹¹. AHSP binds α Hb on the side of the molecule opposite the haem pocket (Fig. 1a). The interface between the FeIII– α Hb and AHSP is nearly identical to that of the FeII– α Hb–AHSP complex, including the preservation of three specific hydrogen bonds and all van der Waals contacts in the centre of the interface⁵. This is remarkable, given that the structure of AHSP-bound α Hb has undergone drastic conformational changes after oxidation.

Although the α Hb–AHSP interface resembles that of the α 1– β 1 complex between α Hb and β Hb (ref. 5) (both involving the hydrophobic surfaces of the G and H helices of α Hb), the orientations of the partner helices are quite different. In the α Hb–AHSP complex, helices α 1 and α 2 of AHSP cross the G and H helices of α Hb at an angle that is about 50° different from that made by the corresponding B and G helices from β Hb (Fig. 1b). This structural difference probably explains why AHSP binding has functional consequences that are completely different from those of β Hb binding.

The haem iron atom is coordinated by both the distal histidine (His 58) and the proximal histidine (His 87) in the FeIII– α Hb–AHSP complex (Fig. 1c). The distances between the haem iron and the N ϵ atom of the imidazole ring are 2.10 and 2.13 Å, respectively, for the distal and the proximal histidine residues. The FeIII–haem group is planar (Fig. 1c). This configuration is in sharp contrast to previously reported structures of erythrocytic haemoglobins, in which haem is bound only by the proximal histidine, and to the FeII– α Hb–AHSP complex, in which haem is coordinated only by the distal histidine and the entire F helix is disordered⁵. In the FeIII– α Hb–AHSP complex, the F helix is well-defined.

Compared with FeII– α Hb bound to AHSP (Fig. 2a) or in HbA (Fig. 2b), the structure of the FeIII– α Hb bound to AHSP shows significant conformational changes involving translocation of main chain atoms by as much as 10 Å (Fig. 2). The root-mean-square-deviation (r.m.s.d.) is 3.2 Å for 135 main-chain C α atoms (residues 2–136) between α Hb in HbA and in the AHSP-bound, oxidized state. This value is in contrast to the r.m.s.d. of about 0.5 Å for the same C α atoms between the T-state and the R-state of α Hb. These conformational changes are more appropriately classified as structural rearrangements because the hydrophobic core is also altered (Fig. 2c).

During the AHSP-induced transition from the FeII to the FeIII state, structural rearrangements in α Hb occur in 80% of its sequence and are concentrated primarily in three segments that are located close to the haem-binding pocket (Fig. 2). The first segment includes the C-terminal half of helix B, the entire C helix, the amino-terminal half of helix E, and the intervening loops (Fig. 2c, left panel). In particular, the distal histidine, His 58, is located in this segment. The second segment encompasses helix F, which harbours the proximal histidine, His 87, and its N-terminal loop (Fig. 2c, middle panel).

¹Department of Molecular Biology, Lewis Thomas Laboratory, Princeton University, Princeton, New Jersey 08544, USA. ²The Children's Hospital of Philadelphia and the University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA. ³School of Molecular and Microbial Biosciences, University of Sydney, NSW 2006, Australia.

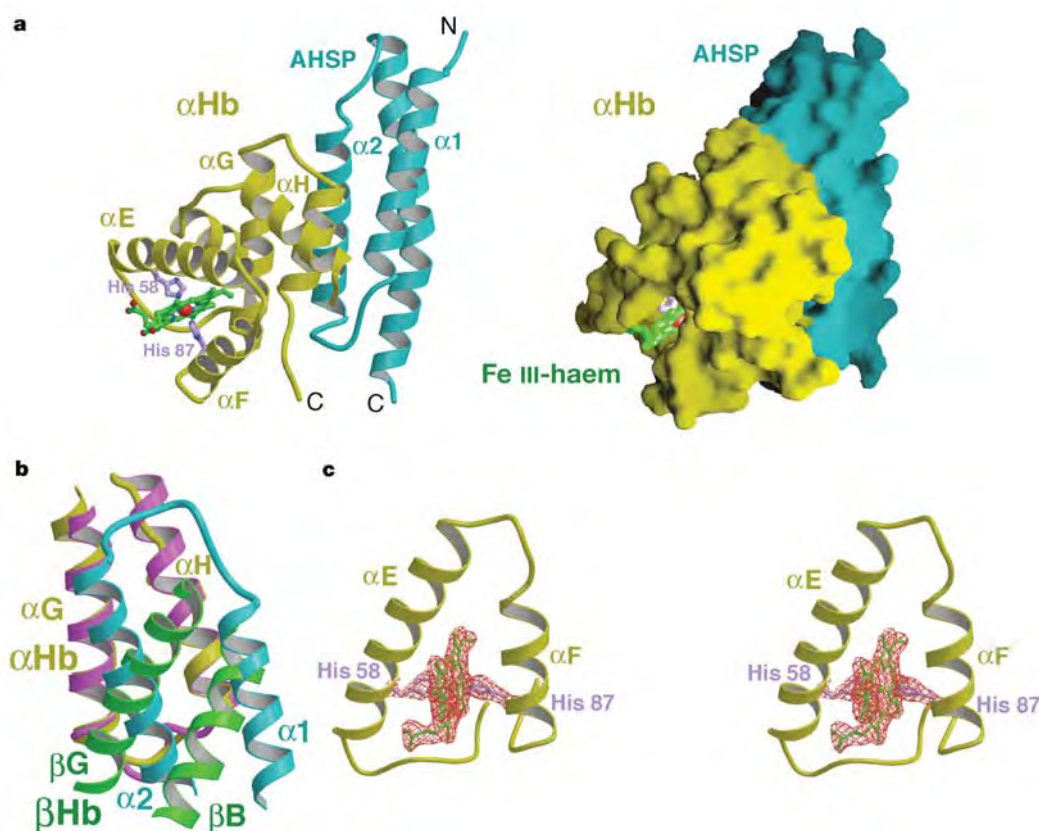


Figure 1 | Structure of the oxidized α Hb bound to AHSP. **a**, Structure of the oxidized α Hb-AHSP complex. In the right panel the structure is represented in a surface model. **b**, Comparison of the α Hb-AHSP interface with that of the α 1- β 1 complex. Superposition was performed on α Hb. α Hb and β Hb in the α 1- β 1 complex are coloured magenta and green, respectively. **c**, A stereo

view of the bis-histidyl configuration on the haem group. The $|F_o - F_c|$ electron-density omit map, contoured at 1.5σ , is shown in red around the haem group and the two histidine residues. The right panel of Figs 1a and 3a were prepared with GRASP²⁹. All other structural figures were made with MOLSCRIPT³⁰.

Structure alteration also occurs at the C-terminal end of α Hb, where the C-terminal half of helix H is transformed into a rigid loop (Fig. 2c, right panel). These changes are accompanied by a significant rearrangement of the side chains that interact with the haem group (Supplementary Fig. 1). In HbA the distance between the N ϵ atom of the imidazole ring of His 58 and haem iron is about 4.3 Å. In the Fe II- α Hb-AHSP complex, this distance is shortened to 2.13 Å because of the coordination of His 58 with the Fe II atom.

The only structural elements that do not undergo significant conformational changes upon the oxidation of α Hb are the entire G helix and the N-terminal half of the H helix (Fig. 2a), which are primarily responsible for interacting with AHSP⁵ (Fig. 1). This feature explains why AHSP can bind specifically to multiple forms of α Hb despite significant conformational differences between them. Perhaps more significantly, the G and H helices of α Hb are also the primary structural elements that interact with β Hb to form the α 1- β 1 complex⁵. This arrangement presumably permits multiple forms of AHSP-bound α Hb to bind β Hb and generate tetrameric HbA.

In the Fe II- α Hb-AHSP complex, the F helix is disordered and the haem surface is open to interaction with solvent⁵. On oxidation of the iron atom, the F helix is reformed to generate the bis-histidyl configuration. The iron atom is much less exposed to solvent, even in comparison with α Hb from HbA (Fig. 3a). Consequently, the ability of the haem iron to act as a catalyst of redox reactions is predicted to decrease. To examine this hypothesis we compared the abilities of Fe II- α Hb-AHSP, Fe III- α Hb-AHSP and free α Hb to participate in redox catalysis.

Haemoglobins undergo a complex set of reactions with the partly reduced oxygen species that occur *in vivo*, which could be influenced

by AHSP-induced structural changes^{12,13}. Both Fe II and Fe III haem react with hydrogen peroxide (H_2O_2) to produce oxyferryl intermediates. The reaction with Fe III haem also generates a free radical within the globin chain^{13,14}. Further reaction of radical oxyferryl haemoglobin with H_2O_2 regenerates Fe III haem and oxygen. Reaction of the non-radical oxyferryl form generates Fe III haem and superoxide¹⁵. The production of superoxide within the haem pocket can result in haem degradation and the release of haem iron and porphyrins¹⁶. We assessed how AHSP affects these processes by comparing H_2O_2 -induced spectrophotometric changes in Fe II- α Hb-AHSP, Fe III- α Hb-AHSP and free α Hb. H_2O_2 induced a rapid decline of Soret absorbance (412 nm) from free α Hb (Fig. 3b, left panel). This decline reflects haem loss, most probably by breakdown as outlined above¹⁶. The initial rate of haem loss was decreased in the Fe III complex but not the Fe II (Fig. 3b, left panel). In later stages of incubation, haem loss was decreased from the Fe II complex.

Table 1 | Refinement statistics

Resolution (Å)	20-2.4
R_{work}/R_{free}	0.222/0.252
Number of atoms	
Protein	3,522
Ligand/ion	86
Water	388
B-factors	
Protein	63.62
Ligand/ion	33.54
Water	67.71
R.m.s. deviations	
Bond lengths (Å)	0.009
Bond angles (°)	1.5

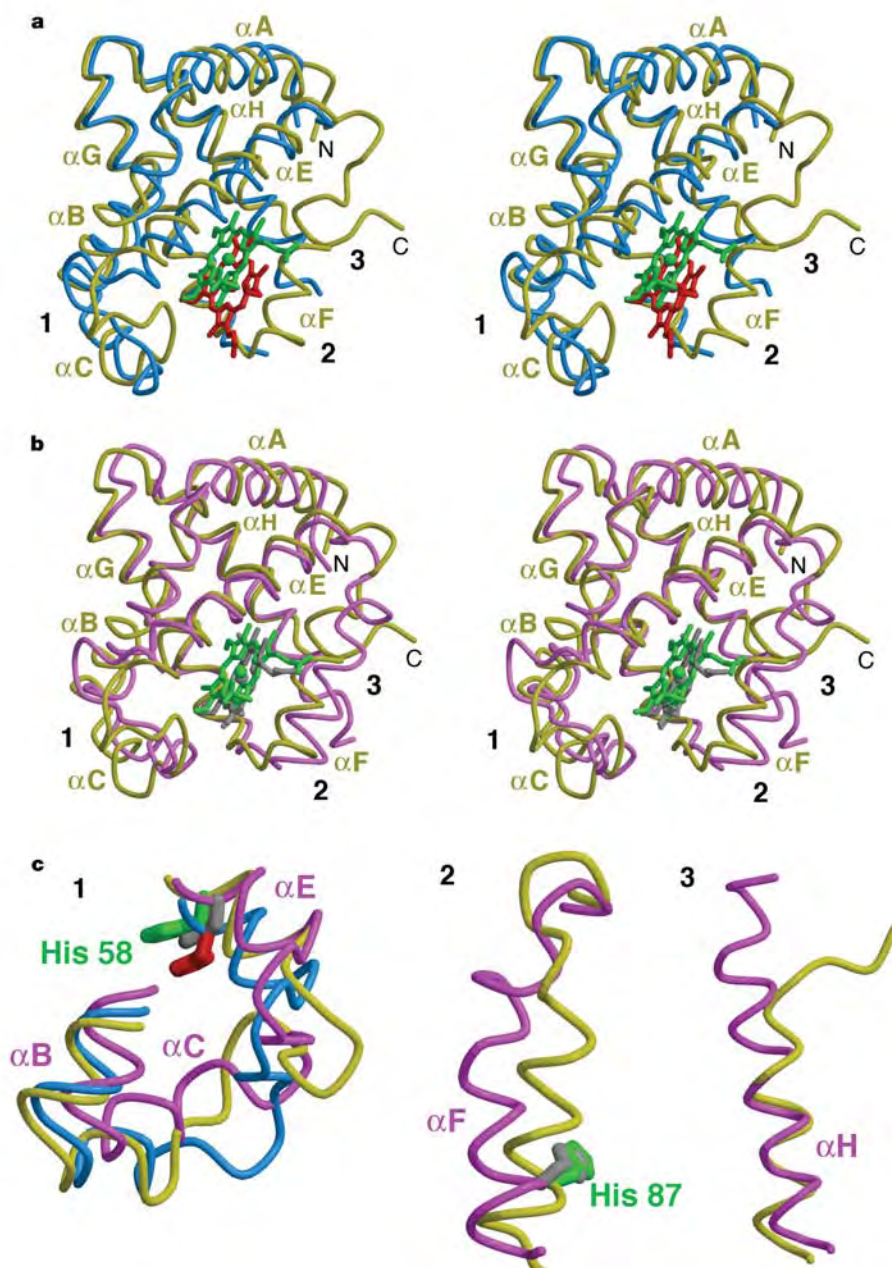


Figure 2 | AHSP-bound oxidized α Hb undergoes structural rearrangements.

a, Stereo-view superposition of the AHSP-bound, Fe III- α Hb (yellow) on the AHSP-bound Fe II- α Hb (blue), with their haem groups in green and red, respectively. Numbers 1–3 mark where significant structural changes occur. **b**, Stereo-view superposition of AHSP-bound Fe III- α Hb (yellow) and α Hb from HbA (magenta), with their haem groups in green and grey, respectively. **c**, A close-up view of the three regions of α Hb that undergo significant structural rearrangements. The proximal and distal histidines are coloured grey, red and green for α Hb from HbA, AHSP-bound Fe II- α Hb and AHSP-bound Fe III- α Hb, respectively. Chains are coloured as follows: magenta, α Hb (reduced, free); cyan, α Hb (reduced, bound to AHSP); yellow, α Hb (oxidized, bound to AHSP).

This was most probably caused by the formation of the more stable Fe III haem state, as indicated by the ultraviolet–visible spectra (Fig. 3b, top right panel).

The bis-histidyl structure of Fe III- α Hb–AHSP could provide protection from peroxide-induced damage by preventing the formation of oxyferryl intermediates. This possibility is supported by an analysis of the changes in α Hb–AHSP spectra (Fig. 3b, right top and bottom panels). Deconvolution of these spectra to known standards indicates that within 1 min, 33% of the haem in the Fe II complex was converted to the ferryl form (Supplementary Fig. 2). Less than 10% of the Fe III haem was converted after 30 min. To confirm these observations, we examined the effect of sulphite addition, which reacts with ferryl haem to form Fe II-sulphHb^{17,18}, which can be distinguished spectrophotometrically¹⁹. From a starting concentration of 18.2 μ M Fe II- α Hb–AHSP, 8.3 $\pm$ 0.05 (s.e.m) μ M sulphHb was formed, whereas only 1.9 $\pm$ 0.07 μ M sulphHb was formed from the Fe III complex. Hence, AHSP-bound Fe II- α Hb interacts with H₂O₂ to produce an oxyferryl intermediate, which after further

reaction produces haem loss/destruction. The formation of ferryl haemoglobin is inhibited after the formation of the bis-histidyl Fe III- α Hb–AHSP complex.

Having established that the Fe III complex decreased oxidant-induced haem damage, we compared the abilities of the complexes to catalyse other redox reactions. The interaction of redox-active haem iron with H₂O₂ generates hydroxyl radicals. We assessed the production of oxidants from the reaction of H₂O₂ (Fig. 3c). Free α Hb produced oxidants at the highest rate, whereas the Fe II complex decreased this rate. The Fe III complex (Fig. 3c, upper panel) further inhibited the production of ROS. The predominant difference in the rate of oxidant generation occurred in the early phase of the reaction (Fig. 3c, lower panel). Our findings show that AHSP-induced conversion of α Hb to the Fe III bis-histidyl complex confers protection from both oxidant-induced haem loss and the generation of free oxidant. Although the Fe II complex had a moderate effect on the preservation of haem, it significantly decreased the production of accessory oxidants in comparison with free α Hb.

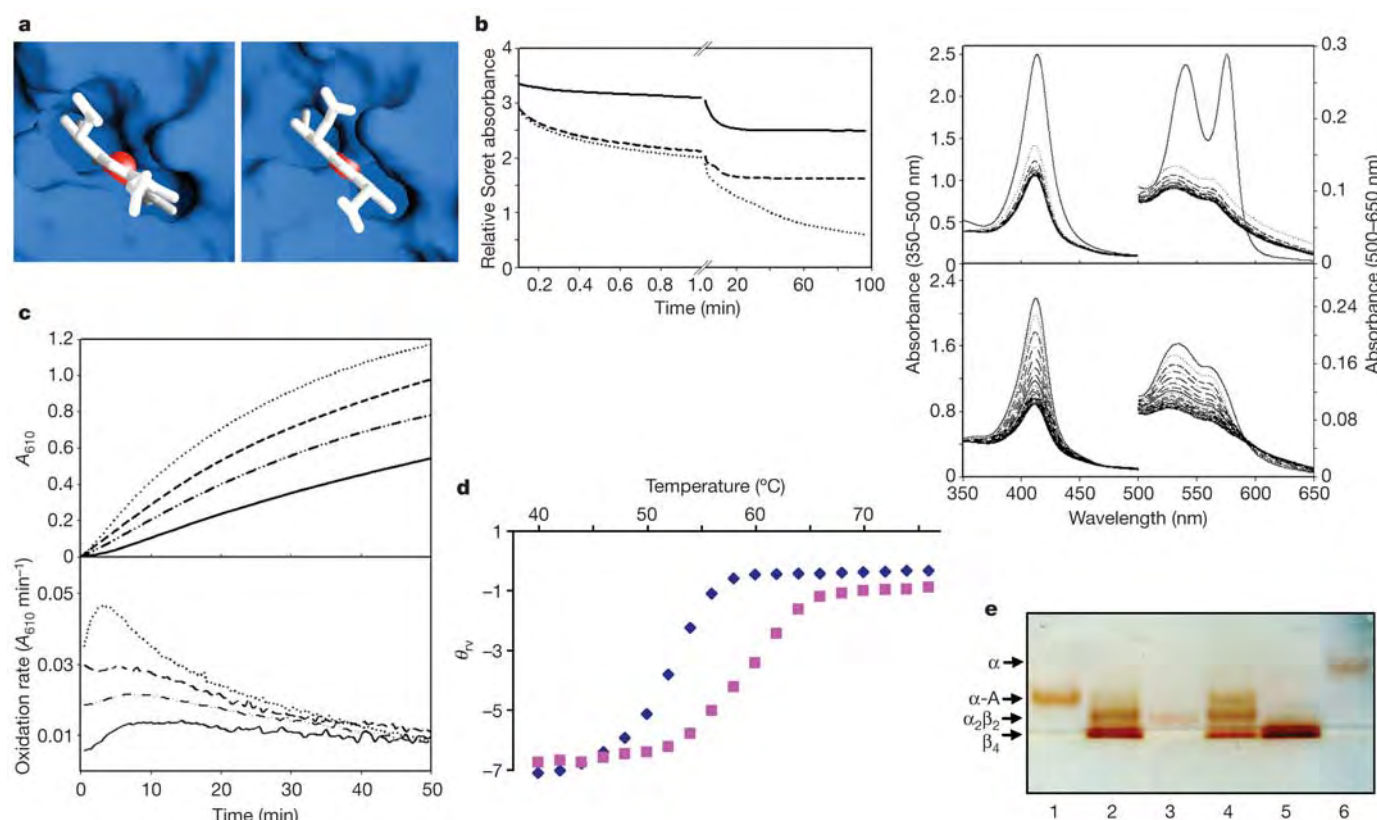


Figure 3 | Protective effects of the bis-histidyl configuration for $\text{Fe III-}\alpha$ Hb.

a, Solvent accessibility is significantly decreased in the haem group of AHSP-bound $\text{Fe III-}\alpha$ Hb (right) in comparison with that for α Hb from HbA (left). **b**, Stability of the haem group in the $\text{Fe III-}\alpha$ Hb-AHSP and $\text{Fe II-}\alpha$ Hb-AHSP complexes after exposure to H_2O_2 . Left, plot of A_{412} against time: dotted line, free α Hb; dashed line, $\text{Fe II-}\alpha$ Hb-AHSP; solid line, $\text{Fe III-}\alpha$ Hb-AHSP. Right top and bottom, complete absorbance spectra of $\text{Fe II-}\alpha$ Hb-AHSP (top) and $\text{Fe III-}\alpha$ Hb-AHSP (bottom) complexes. **c**, Production of secondary oxidants by $\text{Fe III-}\alpha$ Hb-AHSP and $\text{Fe II-}\alpha$ Hb-AHSP complexes. Top, A_{610} was recorded during incubation with H_2O_2 in the presence of TMPD. Bottom, calculated rates of change of A_{610} . Solid

lines, control (H_2O_2 and TMPD alone, in the absence of haem); dotted lines, free α Hb; dashed lines, $\text{Fe II-}\alpha$ Hb-AHSP; dot-dashed lines, $\text{Fe III-}\alpha$ Hb-AHSP. **d**, The melting temperature of α Hb (blue diamonds) is increased by 8°C on binding to AHSP (magenta squares). The melting temperature was calculated by measuring circular dichroism at 222 nm . **e**, The $\text{Fe III-}\alpha$ Hb-AHSP complex can be dissociated by β Hb to form HbA ($\alpha_2\beta_2$). Shown here is a representative native gel-shift assay. The $\text{Fe III-}\alpha$ Hb-AHSP complex, HbA, β Hb and α Hb are in lanes 1, 3, 5 and 6, respectively. The incubation of β Hb with the $\text{Fe III-}\alpha$ Hb-AHSP complex resulted in the formation of HbA (lanes 2 and 4). The amount of β Hb in lane 4 was double that in lane 2. The native gel was directly scanned without staining.

Our present and previous⁵ studies indicate a model in which AHSP protects α Hb. First, through direct binding, AHSP thermodynamically stabilizes α Hb and minimizes the chances that α Hb will denature and thereby form cytotoxic precipitates in cells. In support of this notion, the binding by AHSP increases the melting temperature of α Hb by 8° (Fig. 3d). Second, by facilitating the oxidation of the

haem group and by sequestering the oxidized haem in a redox-inert, bis-histidyl state, AHSP ensures that α Hb does not cause oxidative damage to cells. In addition, in the presence of β Hb, the AHSP-bound, oxidized α Hb can be recruited to form tetrameric Hb (Fig. 3e).

Thus, the crystal structure of the $\text{Fe III-}\alpha$ Hb bound to AHSP

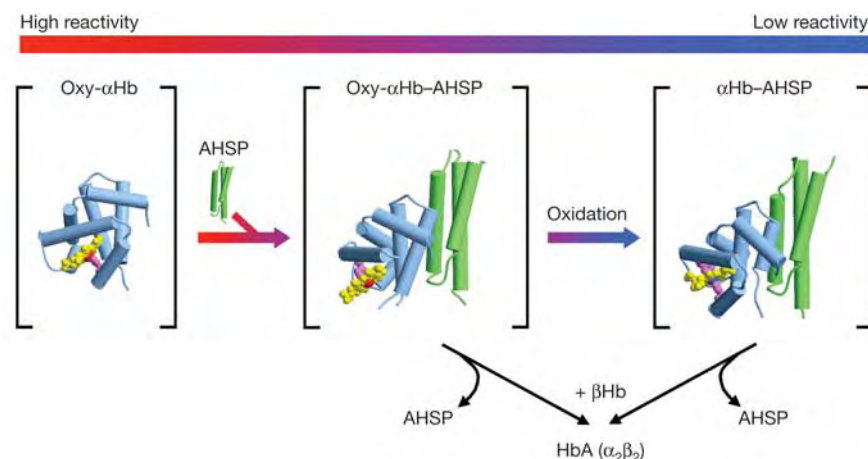


Figure 4 | Mechanisms of AHSP-mediated stabilization of α Hb. Free, oxygenated $\text{Fe II-}\alpha$ Hb (left) exhibits the highest level of chemical reactivity. Binding of AHSP (middle) results in two consequences: decrease in the reactivity of the oxygenated $\text{Fe II-}\alpha$ Hb and rapid formation of the deoxygenated $\text{Fe III-}\alpha$ Hb (right). In the presence of AHSP, the deoxygenated $\text{Fe III-}\alpha$ Hb exhibits the lowest reactivity because of the bis-histidyl configuration at the haem site.

reveals an inert, bis-histidyl configuration that cripples the ability of α Hb to generate ROS and prevents haem loss induced by oxidative stress (Fig. 4). Bis-histidine coordination has previously been identified in the haemoglobin of rice plants²⁰, *Arabidopsis*²¹ and *Drosophila*²² and also within the mammalian haemoglobins cytoglobin and neuroglobin^{23–25}. However, the functions of these bis-histidyl haemoglobins are not fully characterized. Here, formation of the bis-histidyl configuration induced by AHSP binding is an essential step in preventing the cytotoxicity of unpaired α Hb (Fig. 4). These conclusions illustrate a new facet of haemoglobin homeostasis.

METHODS

Protein purification and crystallization. Molecular cloning and protein purification were as described⁵. The purified oxy- α Hb-AHSP complex was oxidized by four molar equivalents of $K_3Fe(CN)_6$ at 0 °C for 5 min. AHSP contained residues 1–91 and the point mutation P30A. Crystals were grown at 4 °C by using the hanging-drop vapour-diffusion method. The well buffer contained 0.1 M MES pH 6.5 and 15% (w/v) PEG2000-monomethyl ether. The crystals belong to the space group C2 and contain two complexes per asymmetric unit. The unit cell has the following dimensions: $a = 65.6$ Å, $b = 113.5$ Å, $c = 79.5$ Å and $\beta = 94.7^\circ$.

Data collection and structure determination. Crystals were equilibrated in buffer containing 0.1 M MES pH 6.5, 15% (w/v) PEG2000-monomethyl ether and 21% (v/v) glycerol, and were flash frozen under a cold nitrogen stream. The native and iron multi-wavelength anomalous dispersion data sets were collected at the NSLS X25 and X12C beamlines, respectively, at the Brookhaven National Laboratories. Data were processed with Denzo and Scalepack²⁶. Positions of the iron atom were identified by using SOLVE²⁷. The atomic models of AHSP and α Hb were refined with the use of CNS²⁸. The final refined model, at 2.4 Å resolution, contains residues 1–91 of AHSP and residues 2–136 of α Hb in both complexes. Circular dichroism spectra on α Hb-AHSP complexes were performed with a π star 180 spectrophotometer (Applied Photophysics).

Native polyacrylamide gel electrophoresis. Oxidized α Hb-AHSP complex (200 μ M) was incubated with equal molar ratio of β Hb on ice for 10 min in 25 μ l of assay buffer containing 20 mM sodium phosphate pH 7.4 and 100 mM NaCl. A 15- μ l sample taken from the mixture was applied to a 6% native polyacrylamide gel under 150 V constant voltage at 4 °C. The electrophoresis buffer contained 25 mM Tris-HCl and 250 mM glycine.

Treatment of α Hb with H_2O_2 . Free α Hb, FeII- α Hb-AHSP and FeIII- α Hb-AHSP complexes were prepared at a haem concentration of 35 μ M. Incubations were performed at 37 °C and initiated by the addition of 350 μ M H_2O_2 . Absorbance spectra were recorded every 5 min. Absorbance at 412 nm (A_{412}) was recorded and is shown against time. Quantification of oxyferryl Hb and sulphHb was performed by linear regression to absorption coefficient spectra for the involved species. SulphHb was generated by the addition of 2 mM Na_2S to FeIII-HbA and excess H_2O_2 (ref. 19).

Production of secondary oxidants by α Hb. FeIII- α Hb-AHSP and FeII- α Hb-AHSP complexes and free α Hb were incubated at room temperature, at a haem concentration of 10 μ M, with H_2O_2 (400 μ M) in the presence of N,N,N',N'-tetramethylbenzene-1,4-diamine (TMPD) in 20 mM sodium phosphate buffer pH 7.4 containing 100 mM NaCl and 10 μ M diethylenetriamine pentaacetic acid (DTPA). A_{610} was recorded every 30 s after the addition of H_2O_2 . For further evaluating the kinetics of oxidant generation, the rates of change of A_{610} were calculated.

Received 7 March; accepted 6 April 2005.

- Bunn, H. F. Subunit assembly of hemoglobin: an important determinant of hematologic phenotype. *Blood* **69**, 1–6 (1987).
- Baglioni, C. Chromosomal and cytoplasmic regulation of haemoglobin synthesis. *Bibl. Haematol.* **29**, 1056–1063 (1966).
- Kihm, A. J. et al. An abundant erythroid protein that stabilizes free α -haemoglobin. *Nature* **417**, 758–763 (2002).
- Kong, Y. et al. Loss of α -hemoglobin-stabilizing protein impairs erythropoiesis and exacerbates β -thalassaemia. *J. Clin. Invest.* **114**, 1457–1466 (2004).
- Feng, L. et al. Molecular mechanism of AHSP-mediated stabilization of α -hemoglobin. *Cell* **119**, 629–640 (2004).
- Scott, M. D. et al. Effect of excess alpha-hemoglobin chains on cellular and membrane oxidation in model beta-thalassemic erythrocytes. *J. Clin. Invest.* **91**, 1706–1712 (1993).
- Nathan, D. G. & Gunn, R. B. Thalassaemia: the consequences of unbalanced hemoglobin synthesis. *Am. J. Med.* **41**, 815–830 (1966).
- Rachmilewitz, E. A. & Schrier, S. L. in *Disorders of Hemoglobin* (eds Steinberg,

- M. H., Forget, B. G., Higgs, D. R. & Nagel, R. L.) 233–251 (Cambridge Univ. Press, Cambridge, 2001).
- Gell, D., Kong, Y., Eaton, S. A., Weiss, M. J. & Mackay, J. P. Biophysical characterization of the α -globin binding protein α -hemoglobin stabilizing protein. *J. Biol. Chem.* **277**, 40602–40609 (2002).
- Santiveri, C. M. et al. NMR structure of the α -hemoglobin chaperone AHSP: Insights into conformational heterogeneity and binding. *J. Biol. Chem.* **279**, 34963–34970 (2004).
- Dickerson, R. E. & Geis, I. *Hemoglobin: Structure, Function, Evolution and Pathology* (Benjamin/Cummings, Menlo Park, 1983).
- Nagababu, E. & Rifkind, J. M. Heme degradation by reactive oxygen species. *Antioxid. Redox Signal.* **6**, 967–978 (2004).
- Rifkind, J. M., Ramasamy, S., Manoharan, P. T., Nagababu, E. & Mohanty, J. G. Redox reactions of hemoglobin. *Antioxid. Redox Signal.* **6**, 657–666 (2004).
- Alayash, A. I., Ryan, B. A., Eich, R. F., Olson, J. S. & Cashion, R. E. Reactions of sperm whale myoglobin with hydrogen peroxide. Effects of distal pocket mutations on the formation and stability of the ferryl intermediate. *J. Biol. Chem.* **274**, 2029–2037 (1999).
- Nagababu, E. & Rifkind, J. M. Reaction of hydrogen peroxide with ferrylhemoglobin: superoxide production and heme degradation. *Biochemistry* **39**, 12503–12511 (2000).
- Reeder, B. J., Svistunenko, D. A., Sharpe, M. A. & Wilson, M. T. Characteristics and mechanism of formation of peroxide-induced heme to protein cross-linking in myoglobin. *Biochemistry* **41**, 367–375 (2002).
- Berzofsky, J. A., Peisach, J. & Horecker, B. L. Sulfheme proteins. IV. The stoichiometry of sulfur incorporation and the isolation of sulfhemin, the prosthetic group of sulfmyoglobin. *J. Biol. Chem.* **247**, 3783–3791 (1972).
- Herold, S. & Rehmann, F. J. Kinetic and mechanistic studies of the reactions of nitrogen monoxide and nitrite with ferryl myoglobin. *J. Biol. Inorg. Chem.* **6**, 543–555 (2001).
- Berzofsky, J. A., Peisach, J. & Blumberg, W. E. Sulfheme proteins. I. Optical and magnetic properties of sulfmyoglobin and its derivatives. *J. Biol. Chem.* **246**, 3367–3377 (1971).
- Arredondo-Peter, R. et al. Rice hemoglobins. Gene cloning, analysis, and O_2 -binding kinetics of a recombinant protein synthesized in *Escherichia coli*. *Plant Physiol.* **115**, 1259–1266 (1997).
- Perazzolli, M. et al. *Arabidopsis* nonsymbiotic hemoglobin AHb1 modulates nitric oxide bioactivity. *Plant Cell* **16**, 2785–2794 (2004).
- Hankeln, T. et al. Characterization of *Drosophila* hemoglobin. Evidence for hemoglobin-mediated respiration in insects. *J. Biol. Chem.* **277**, 29012–29017 (2002).
- Burmester, T., Weich, B., Reinhardt, S. & Hankeln, T. A vertebrate globin expressed in the brain. *Nature* **407**, 520–523 (2000).
- Sugimoto, H. et al. Structural basis of human cytoglobin for ligand binding. *J. Mol. Biol.* **339**, 873–885 (2004).
- Svistunenko, D. A. et al. The pH dependence of naturally occurring low-spin forms of methaemoglobin and metmyoglobin: an EPR study. *Biochem. J.* **351**, 595–605 (2000).
- Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276**, 307–326 (1997).
- Terwilliger, T. C. & Berendzen, J. Correlated phasing of multiple isomorphous replacement data. *Acta Crystallogr. D* **52**, 749–757 (1996).
- Brunger, A. T. et al. Crystallography and NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr. D* **54**, 905–921 (1998).
- Nicholls, A., Sharp, K. A. & Honig, B. Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons. *Proteins Struct. Funct. Genet.* **11**, 281–296 (1991).
- Kraulis, P. J. Molscript: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallogr.* **24**, 946–950 (1991).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank M. Becker and A. Sexana at BNL NSLS beamlines for help. This work was supported by grants from the National Institutes of Health (to Y.S. and M.J.W.), the American Heart Association (to A.J.G.) and the Cooley's Anemia Foundation (S.Z.).

Author Contributions L.F., S.Z. and L.G. contributed equally to this work. L.F. crystallized the oxidized α -haemoglobin bound to AHSP. L.F. and L.G. solved the structure. S.Z. and L.F. performed the biochemical experiments in Fig. 3. Y.S. and A.J.G. wrote the paper. All authors discussed the results and commented on the manuscript.

Author Information The atomic coordinates of the oxidized α Hb-AHSP complex have been deposited in the Protein Data Bank with the accession number 1Z8U. Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to A.J.G. (gow@email.chop.edu) or Y.S. (yshi@molbio.princeton.edu).

LETTERS

Structural characterization of the molecular platform for type III secretion system assembly

Calvin K. Yip¹, Tyler G. Kimbrough², Heather B. Felise³, Marija Vuckovic¹, Nikhil A. Thomas⁴, Richard A. Pfuetzner¹, Elizabeth A. Frey¹, B. Brett Finlay⁴, Samuel I. Miller^{2,3} & Natalie C. J. Strynadka¹

Type III secretion systems (TTSSs) are multi-protein macromolecular 'machines' that have a central function in the virulence of many Gram-negative pathogens by directly mediating the secretion and translocation of bacterial proteins (termed effectors) into the cytoplasm of eukaryotic cells¹. Most of the 20 unique structural components constituting this secretion apparatus are highly conserved among animal and plant pathogens and are also evolutionarily related to proteins in the flagellar-specific export system. Recent electron microscopy experiments have revealed the gross 'needle-shaped' morphology of the TTSS^{2–4}, yet a detailed understanding of the structural characteristics and organization of these protein components within the bacterial membranes is lacking. Here we report the 1.8-Å crystal structure of EscJ from enteropathogenic *Escherichia coli* (EPEC), a member of the YscJ/PrgK family whose oligomerization represents one of the earliest events in TTSS assembly⁵. Crystal packing analysis and molecular modelling indicate that EscJ could form a large 24-subunit 'ring' superstructure with extensive grooves, ridges and electrostatic features. Electron microscopy, labelling and mass spectrometry studies on the orthologous *Salmonella typhimurium* PrgK within the context of the assembled TTSS support the stoichiometry, membrane association and surface accessibility of the modelled ring. We propose that the YscJ/PrgK protein family functions as an essential molecular platform for TTSS assembly.

Gram-negative bacteria use TTSSs to assemble virulence-associated organelles on their cell surface, a process essential to their pathogenesis. TTSSs transport protein substrates across both bacterial membranes in a single-step process that is independent of the Sec pathway and does not involve a periplasmic intermediate. The macromolecular structures of these virulence systems are composed of two distinct parts: the needle complex (NC) and the translocon. The NC, which has been purified and examined by electron microscopy, spans the bacterial envelope and resembles a molecular syringe^{3,6}. The translocon proteins, in contrast, are secreted through the NC and associate with the host cell membrane, where they function to transfer proteins into the cytosol of the host cell. The translocated proteins alter basic cell functions such as signal transduction, cytoskeletal architecture, membrane trafficking, cytokine gene expression and cell death⁷.

Although significant progress has been made in the identification and characterization of the translocated proteins, the detailed structural characteristics and precise molecular organization of the TTSS remain poorly understood. Recent genetic and biochemical studies of the prototypical *S. typhimurium* SPI-1 TTSS revealed that assembly of this macromolecular complex initiates with a *sec*-dependent phase involving the localization and oligomerization of

the following three proteins within the bacterial membranes: InvG, PrgK and PrgH^{6,8}. InvG belongs to the well-characterized secretin superfamily of proteins, which are also involved in type II secretion, filamentous phage assembly and type IV fimbrial morphogenesis. Like other secretins, InvG forms a large homomultimeric annular complex in the outer membrane, which putatively mediates the passage of protein substrates⁹. In contrast, PrgK and PrgH are believed to associate in the inner membrane to form the base of the NC⁵. Recent cryo-electron microscopy and reconstruction studies have revealed a variable 19–22-fold symmetry at the base of the *S. typhimurium* NC, but the identity of the proteins responsible for generating this symmetry could not be deduced². Although little is known about PrgH, and orthologues for this protein have not yet been identified in all species, PrgK belongs to the highly conserved YscJ/PrgK protein family and is predicted to be periplasmic and anchored to the bacterial membrane through its amino-terminal

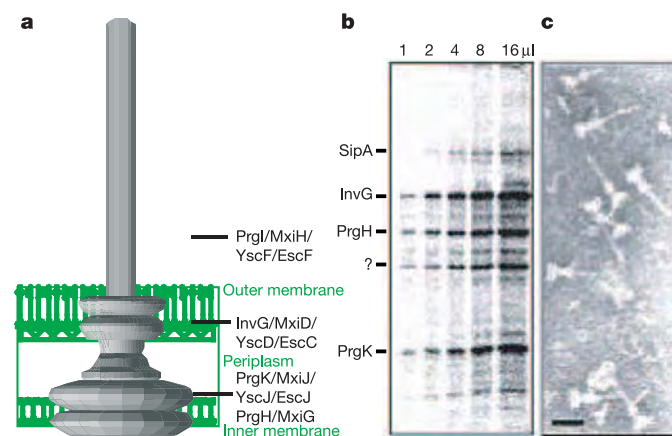


Figure 1 | Stoichiometric analysis of *S. typhimurium* NC base components. **a**, Diagram of bacterial TTSS. The three components of the base and the needle component of four representative species are labelled with respect to their putative locations in the bacterial envelope in the following order: *Salmonella*, *Shigella*, *Yersinia* and EPEC. **b**, Autoradiograph of [³⁵S]Met/Cys-labelled needle complexes resolved on a 10% SDS-PAGE gel and imaged with the Storm 840 PhosphorImager. A range of dilutions were loaded as shown. The needle complex components InvG, PrgH and PrgK are indicated, as well as two co-purifying components, SipA and an unknown fragment. **c**, Electron micrograph of the needle complex preparation used for stoichiometric analysis. Scale bar, 50 nm.

¹Department of Biochemistry and Molecular Biology, University of British Columbia, 2146 Health Sciences Mall, Vancouver, British Columbia, Canada V6T 1Z3. ²University of Washington, Department of Molecular and Cellular Biology, Genome Sciences, HSB K-140, Box 357710, Seattle, Washington 98195, USA. ³Department of Microbiology and Medicine, HSB K-140, Box 357710, Seattle, Washington 98195, USA. ⁴Michael Smith Laboratories, University of British Columbia, 301-2185 East Mall, Vancouver, British Columbia, Canada V6T 1Z4.

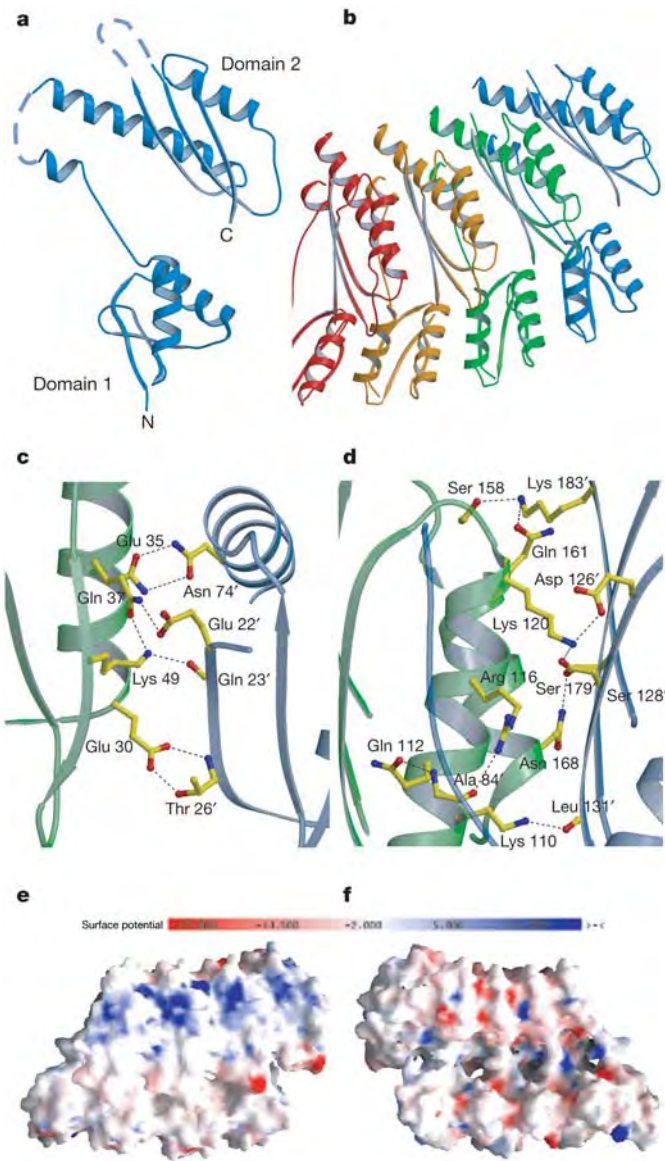


Figure 2 | EscJ structure and intermolecular interactions. **a**, Ribbon representation of EscJ. The triangular monomer consists of two topologically similar mixed α/β domains. **b**, Arc-shaped EscJ tetramer in the crystallographic asymmetric unit. The four molecules are viewed with their N termini projecting into the page. **c**, Interface between domain 1 of two EscJ monomers. The two monomers are coloured green and blue, and the residues involved in hydrogen-bonding interactions are highlighted with dotted lines depicting potential hydrogen bonds. **d**, Interface between domain 2 of two EscJ monomers. Residues from two α -helices of the green monomer interact with residues from the β -sheet of the blue monomer. **e**, Surface representation of tetrameric EscJ coloured by electrostatic potential. The orientation of the molecules is approximately the same as in **b**. Positively charged patches are observed near the top of the outer face of the tetramer. **f**, Molecular surface of tetrameric EscJ on the inner face. The surface shown in **e** is rotated by 180° and shown with the same colouring schemes.

Table 1 | Stoichiometry of proteins of the NC base component

Proteins	Molecular mass (kDa)	N-terminal sequence	Processing event	Sulphur number	Stoichiometry	
					Relative	Estimated subunit number
InvG	62	MX ₂₃ ↑ SEK...	Signal peptide	14	1.0 ± 0.01*	13.0 ± 1.0
PrgH	50	METSK...	None	7	1.3 ± 0.02*	17.4 ± 1.4
PrgK	28	MX ₁₂ LAG ↑ CK...	Signal peptide	5	1.7 ± 0.03*	22.0 ± 1.7

*These values are the weighted averages of four independent experiments; errors are s.d. Statistical analysis indicated that the observed differences in mean intensities were significant (two-factor analysis of variance, $F = 18.4$, $P < 0.0001$).

lipid and possibly a hydrophobic carboxy-terminal transmembrane segment⁴. Members of this family share sequence similarity with the central domain of the flagellar FlhF protein, which forms the MS (membrane/supramembrane)-ring in the inner membrane and initiates assembly of the flagellar-specific secretion system¹⁰. Despite the crucial roles of PrgK and FlhF in the morphogenesis and function of their respective secretion systems, high-resolution structural information for these proteins is unavailable.

To characterize the molecular organization of the TTSS structure we purified NCs from *S. typhimurium* grown in the presence of [³⁵S]methionine and [³⁵S]cysteine, and analysed them by autoradiography to determine the subunit stoichiometry of three structural components, namely InvG, PrgH and PrgK (Fig. 1). After normalization for sample loading and sulphur content of the proteins, the relative ratios of the InvG, PrgH and PrgK proteins were estimated to be 1.0, 1.3 and 1.7, respectively (Table 1). Previous electron microscopy studies have shown that InvG is identical in size to other secretins and that the outer-membrane oligomeric structures formed by these proteins consist of 12–14 subunits^{9,11–13}. Assuming that the number of InvG subunits in the outer membrane ring falls within this narrow range (13 ± 1.0 ; errors are s.d.), the approximate numbers of PrgH and PrgK molecules per NC are calculated to be 17.4 ± 1.4 and 22.0 ± 1.7 , respectively. The estimated stoichiometry of PrgK is similar to the number of FlhF molecules (26) within the MS-ring of the flagellar export system as determined with similar methods¹⁴.

To gain further insights into the structural properties and the intermolecular interactions that stabilize the oligomeric YscJ/PrgK proteins, we have determined the 1.8-Å crystal structure of EscJ from EPEC, a protein that has the advantage of lacking the predicted C-terminal transmembrane segment present in many members of this family including PrgK (Supplementary Table 1 and Supplementary Fig. 1). EscJ crystallized in the space group $P6_5$, with four molecules in the crystallographic asymmetric unit. The refined model shows that the EscJ monomer is a flat triangular molecule consisting of two mixed α/β domains with similar and unusual topologies (Fig. 2a). These two globular regions are connected by a relatively extended linker region (Gly 77 to Ala 84) that might serve to fine-tune the angular orientation of the two domains. The four EscJ molecules in the asymmetric unit adopt highly similar conformations, with the matched 153 C α atoms superimposing with root-mean-squared deviation (r.m.s.d.) values in the range 0.5–1.0 Å. The four monomers, which are positioned in tandem, form a relatively compact arc-shaped tetramer (Fig. 2b).

Several notable features were observed when examining the interactions between the EscJ monomers. First, each EscJ monomer is involved in extensive interactions with its two neighbouring molecules over its entire length, resulting in the burial of 35% (about 3,500 Å²) of the total solvent-accessible surface. Second, the EscJ–EscJ interface contains a significant number of charged residues, and several of them are involved in hydrogen-bonding interactions. Most of these interactions are repeated between any two monomers. More specifically, the side chains of Glu 30 and Lys 49 in domain 1 form hydrogen bonds with the backbone amide and carbonyl of residues Thr 26' and Gln 23' of the adjacent chain, and the side chains of Glu 35 and Gln 37 form hydrogen bonds to the side chains of Asn 74' and Glu 22', respectively (Fig. 2c). Similarly, the

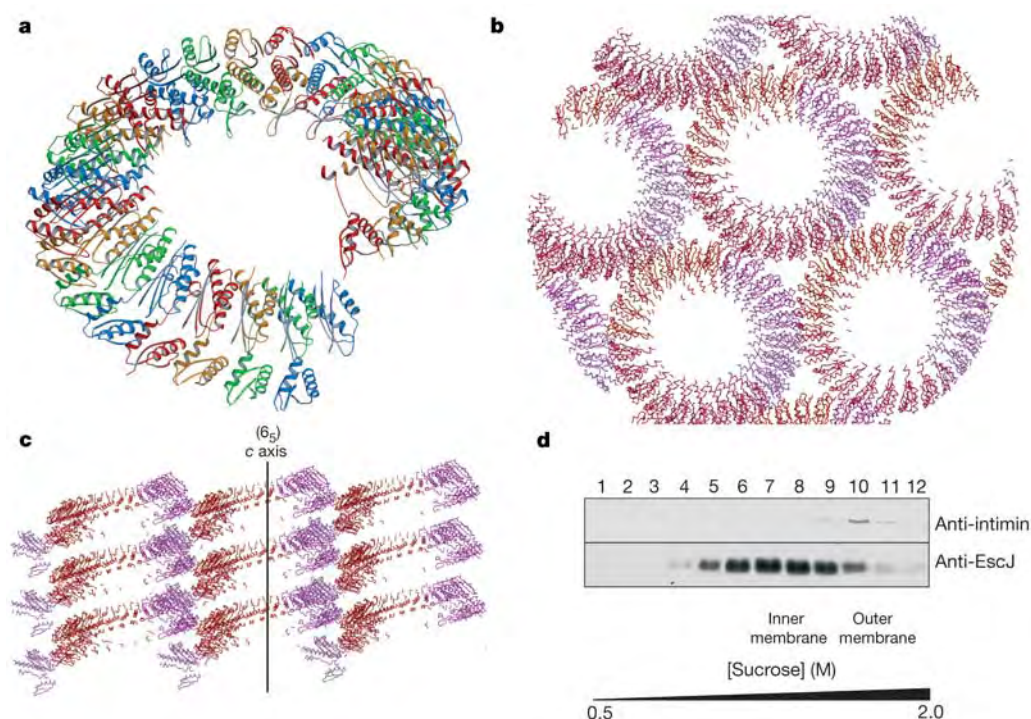


Figure 3 | EscJ forms a superhelix in the crystal but is anchored to the inner membrane *in vivo*. **a**, The EscJ tetramer generates a superhelix by repeating itself six times in a circular fashion along the crystallographic 6_5 screw axis. A ribbon representation of one full turn of the superhelix, which contains 24 EscJ molecules, is shown. **b**, Molecular packing of EscJ viewed above the 6_5 screw axis. For clarity only the $C\alpha$ backbone of the molecules is shown. The superhelical arrangement extends along the entire length of the crystal, resulting in long solvent-filled central channels. **c**, Molecular packing of EscJ viewed in projection along the 6_5 axis of the crystal. The superhelices in this view resemble stacks of tilted rings. **d**, Total membranes were extracted from EPEC and fractionated on a sucrose-density gradient. As shown from the western blotting of the fractions, EscJ localizes to the inner membrane, whereas the bacterial surface adhesin, intimin, is found exclusively in the outer membrane.

side chains of Lys 110, Arg 116, and Gln 112 of the slightly larger domain 2 make hydrogen bond contacts with the backbone carbonyls and amide of Leu 131' and Ala 84', respectively, and the side chains of Lys 120 of helix $\alpha 3$ and Ser 158, Gln 161 and Asn 168 of helix $\alpha 4$ form hydrogen bonds with residues Asp 126', Ser 128', Lys 183' and Ser 179' of the three-stranded β -sheet in the adjacent chain (Fig. 2d). A consequence of the burial of charged residues and

formation of these hydrogen bonds at the oligomeric interface is a marked overall change in the disposition of electrostatic charge on the molecular surface. Although charges seem to be scattered randomly along the surface of the EscJ monomer, two charged patches (one positive and one negative) accumulate on opposite faces (referred to as the outer face and inner face) of the EscJ tetramer (Fig. 2e, f). Third, the tetramer surface at the inner face consists of

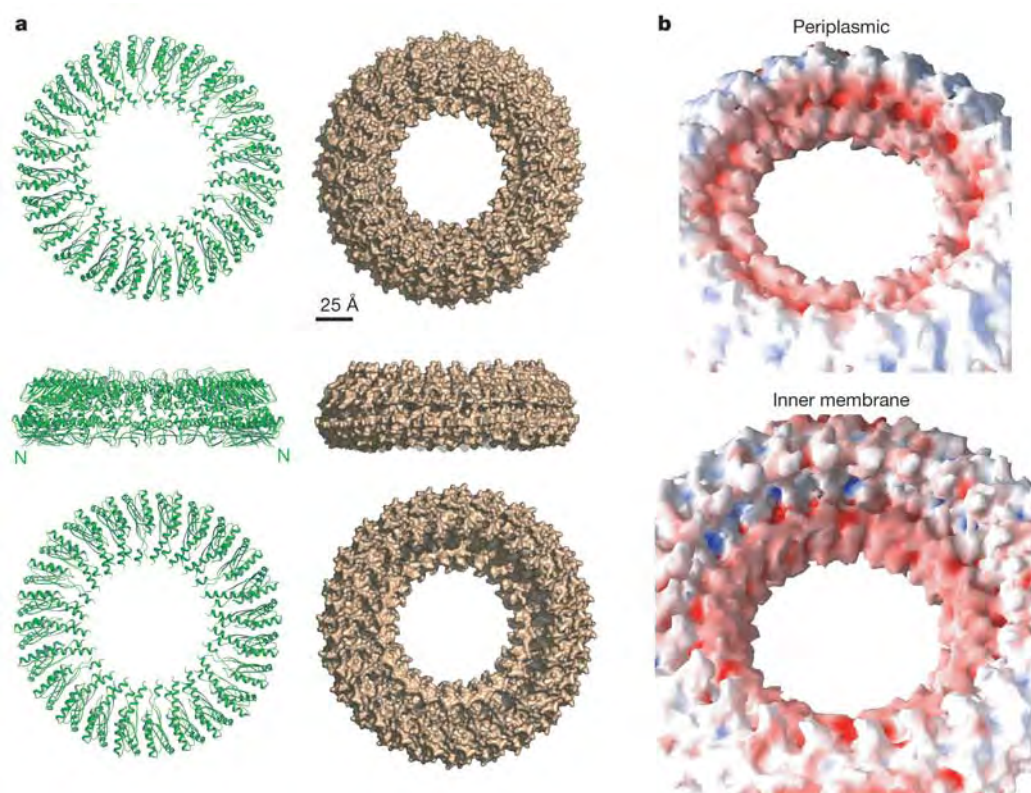


Figure 4 | Modelling and surface electrostatic analysis of the EscJ ring. **a**, Ribbon and surface representation of the modelled 24-subunit EscJ ring. The side view shows the two-layered exterior structure, which is closely analogous to the FlIF ring structure determined from cryoEM studies. The N termini for all subunits are located at the wide face of the ring, with two of them labelled N in the ribbon diagrams. **b**, Surface electrostatics of the EscJ ring. The periplasmic face and the inner-membrane face of the channel are shown, and the ring is slightly tilted to allow a better visualization of the depth. The trench region surrounding the central channel at the periplasmic face is quite negatively charged. The spacious interior of the ring shows the same set of cavities as those observed in the tetramer.

four deep cavities lined with several lysine and arginine residues (Fig. 2f). Last, in addition to the hydrogen-bonding interactions, there are many van der Waals and hydrophobic contacts dispersed across the intermolecular interface. The extensive and specific nature of these intermolecular contacts indicates that oligomerization is an intrinsic property of EscJ and its orthologues.

Further analysis of the molecular packing in the EscJ crystal revealed that symmetry-related tetramers pack into a superhelical structure by means of similar interactions to those detailed above (Fig. 3a), indicating that EscJ might form a much larger assembly. This superstructure, whose helical axis is parallel to the crystallographic 6_5 screw axis, contains 24 EscJ monomers per helical turn

with a pitch equal to the length of the crystallographic c axis ($67 \text{ \AA}$). The centre of this superhelix, with an estimated width of $75 \text{ \AA}$, is solvent filled and completely devoid of protein (Fig. 3b). The N termini of all the EscJ molecules in the crystal project in the same direction with respect to the superhelix axis, and when viewed along this axis the superstructure resembles a stack of tilted rings (Fig. 3c).

Members of the YscJ/PrgK family have been shown to be lipidated at their N termini⁴ (Supplementary Fig. 2), and our sucrose gradient experiments on native EscJ established that the acylated amino termini of these proteins localize to the bacterial inner membrane (Fig. 3d). The non-lipidated nature of the protein we used for crystallization and the absence of a membrane in the crystal alleviate

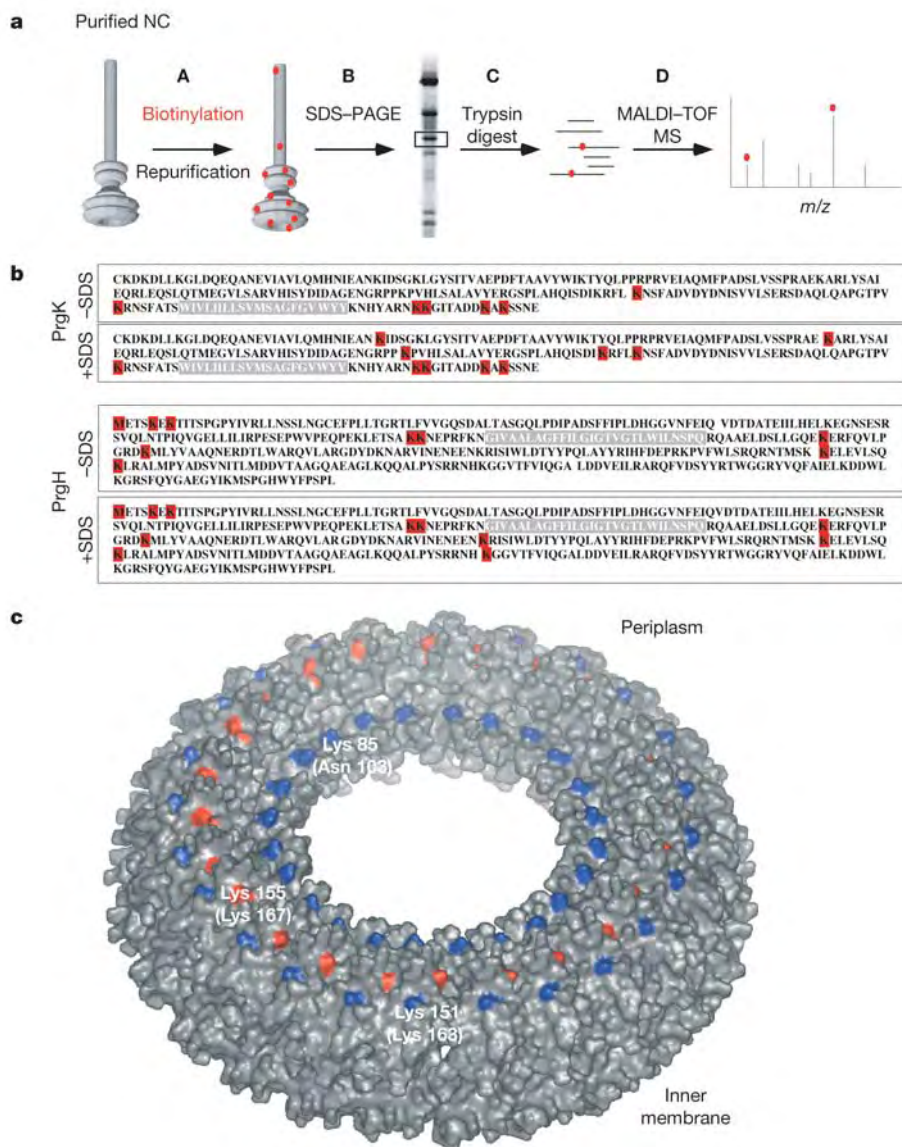


Figure 5 | Surface mapping of *S. typhimurium* NC with limited biotinylation and MALDI-TOF MS. **a**, Diagram of NC surface mapping. Purified NCs were treated with the biotinylation reagent. Protein components of biotinylated NCs were then separated by SDS-PAGE and revealed by Coomassie Blue staining. Biotinylated protein bands corresponding to PrgH, PrgK and InvG proteins were excised individually and digested *in situ* with trypsin. Tryptic peptide fragments were eluted and their masses determined by MALDI-TOF MS. The resulting mass spectra were then analysed with the ExPASy proteomics tool (FindPept program, (<http://www.expasy.ch/tools/findpept.html>)) to identify tryptic peptides containing biotinylated lysine residues. **b**, Graphical representation of MALDI-TOF MS peptide mapping

analysis of biotinylated peptide fragments from PrgK after *in situ* trypsin digestion and streptavidin affinity purification (Supplementary Table 2). Purified NC was either labelled with biotin directly (-SDS) or was denatured by being boiled in 2% SDS (+SDS) before being labelled. Biotin-modified lysine residues are highlighted in red. Residues corresponding to the single transmembrane segment of PrgK are highlighted in grey. **c**, Surface lysines of PrgK are inaccessible to biotinylation. Three surface-exposed lysines of PrgK are mapped on a surface representation of the EscJ ring model in grey, with the corresponding residues in EscJ labelled in parentheses. Inaccessible residues are coloured blue; accessible or biotinylated residues are shown in red.

this N-terminal lipid anchoring, which in the physiological environment would constrain the EscJ oligomeric structure to be in a plane. We have therefore modelled EscJ as a flat ring based on the crystallographic superhelical structure. The modelling procedure, which involves a slight translation of the EscJ monomers along the crystallographic 6₅ screw axis, projects one turn of the superhelix into a two-dimensional plane and results in little change in the directionality and extensive intermolecular interaction of the individual molecules.

The resulting EscJ ring maintains the 24-fold symmetry observed in the crystal, a number in keeping with the estimated stoichiometry of PrgK (22.0 ± 1.7) in the *S. typhimurium* NC (Table 1). The refined EscJ ring model has an overall diameter of 180 Å and height of 52 Å (Fig. 4), matching the values previously estimated for the inner membrane ring from electron microscope images of purified EPEC NCs¹⁵. The vast majority of intermolecular contacts observed in the crystal structure are also achieved in the modelled ring. Further support of this model comes from limited biotinylation–mass spectrometry surface mapping experiments of purified *S. typhimurium* NC. The data showed that the lysine residues of PrgK that are predicted to be buried at the intermolecular interfaces (Lys 4, Lys 8, Lys 31 and Lys 56), on the basis of the EscJ ring model, are indeed not modified (Fig. 5 and Supplementary Table 2).

Even though we cannot definitely conclude that EscJ and its orthologues strictly adopt an oligomerization number of 24 *in vivo*, we believe the extensive complimentary intermolecular contacts and the overall electrostatic and shape properties observed in our model would be largely maintained in any energetically favourable physiological oligomer. The N termini of all EscJ subunits localize to the wider face of the supermolecular complex, indicating that the ring might be anchored to the inner membrane at this end with the opposite face extending towards the periplasm (Fig. 4). The central channel of the ring constricts from about 120 Å at the membrane face to about 73 Å at the periplasmic face. The interior is characterized by a prominent negatively charged ‘ridge’ and is fully circumscribed by the unusual positively charged cavities observed in the tetrameric structure (Fig. 4). In contrast, the periplasmic opening of the central channel is surrounded by a deep and negatively charged ‘trench’, about 32 Å in height and 11 Å in width (Fig. 4). The overall dome-shaped morphology and the two-layered exterior appearance of the EscJ ring are strikingly similar to the recent 22-Å cryo-electron microscopy structure of the flagellar FlrF ring¹⁶. Thus, the domain of FlrF mediating MS-ring formation probably lies in regions conserved with the PrgK family of proteins (see Supplementary Fig. 1), whereas residues present in FlrF but not in the PrgK family might be responsible for forming the rod-like extension that protrudes from the MS-ring into the bacterial periplasm¹⁶.

Our structural analysis of the EscJ ring not only presents an atomic snapshot of one of the earliest structures generated in the TTSS assembly process, but also reveals features indicative of a role as the molecular platform for subsequent construction of the secretion apparatus. First, the lack of transmembrane segments in the individual subunits means that the oligomeric EscJ/PrgK ring is positioned on top of rather than within the inner membrane and therefore is not likely to generate a hollow pore. The specific patch of membrane where the ring localizes might function to recruit critical transmembrane components such as EscR/SpaP, EscS/SpaQ, EscT/SpaR, EscU/SpaS and EscV/InvA (or their equivalents in other species). The large interior together with the regularly spaced and highly polarized electrostatic features provide sufficient room and specific interaction sites for anchoring these transmembrane components, many of which are predicted to contain periplasmic domains that might interact with the EscJ/PrgK ring. Second, the prominent negatively charged trench in the periplasmic opening of the EscJ/PrgK ring, which the PrgK biotinylation study indicates to be inaccessible (Fig. 5), might serve as the critical adaptor region for binding the central rod assembly consisting of EscF/PrgL–PrgJ. These proteins have been proposed to form a helical polymeric structure

that connects the inner and outer membrane rings. Last, the greater degree of biotin modification that we have observed for PrgH in comparison with PrgK (Fig. 5 and Supplementary Table 3) and the inaccessibility of the predicted outer rim of the PrgK ring in the intact NC support our working model in which the surface of the PrgK oligomer is encompassed by the PrgH polymer⁵.

Collectively, these biochemical and structural analyses of PrgK and EscJ have provided new insights into the initial stages of TTSS assembly and probably support a critical role of the YscJ/PrgK family of proteins as an inner-membrane-associated molecular platform for TTSS assembly.

METHODS

Stoichiometric analysis of *S. typhimurium* NC components. The NC-overproducing strain TK385 (ref. 6) was grown in medium containing 5 mCi of EXPRE³⁵S [³⁵S]Protein Labelling Mix (PerkinElmer). Radiolabelled NCs were separated by SDS–PAGE. The gels were fixed, dried, and autoradiographed with a Storm 840 PhosphorImager (Molecular Dynamics). The data were analysed with the Imagequant V1.2 software package (Molecular Dynamics). The intensity data set ($N = 60$; five measurements for three proteins on four runs) was analysed with a two-factor analysis of variance that included protein identity (InvG, PrgH and PrgK) and run number (runs 1–4) as fixed effects.

Surface accessibility analysis of *S. typhimurium* NC components with limited biotinylation and matrix-assisted laser desorption/ionization–time-of-flight (MALDI–TOF) MS. The biotinylation reagent EZ-Link Sulfo-NHS-LC-Biotin (Pierce) was added to purified NC and the reaction was allowed to proceed for 2 h. After being quenched with Tris–HCl, samples were separated by SDS–PAGE and the gels were stained with Coomassie Brilliant Blue R-250. Polyacrylamide gel slices were fragmented, destained during incubation overnight in 50% (v/v) methanol, and dehydrated by incubation in acetonitrile. The gel fragments were subsequently rehydrated in trypsin solution (Promega). After incubation for 30 min on ice, excess trypsin was removed and ammonium bicarbonate was added. The resulting peptides were eluted from the gel with several changes of extraction buffer containing 50% acetonitrile and 5% formic acid before being dried by evaporation. Peptides were solubilized and purified with streptavidin–affinity columns. Peptides were further purified by C₁₈ Zip Tips (Millipore). Samples of purified peptides (0.6 µl) were mixed on a target with 0.6 µl of alpha-cyano-4-hydroxycinnamic acid (CHCA) in acetonitrile and left to dry. MALDI spectra were acquired with a Bruker BIFLEX III spectrometer (Bruker Daltonics) in reflection mode.

Membrane localization of EscJ. Total membranes were first prepared from EPEC grown in DMEM medium, and separation of inner and outer membranes was performed with sucrose gradients from 0.5 to 2.0 M as described elsewhere¹⁷. **Expression and purification of EscJ.** *E. coli* BL21(ΔDE3) was transformed with the respective pET-41EscJ(21–190) constructs, grown at 37 °C to a D_{600} of 0.6–0.8, and induced overnight with 0.5 mM isopropyl β-D-thiogalactoside at 20 °C. Cultures were harvested and cells were lysed with a pressurized homogenizer (Avestin). The soluble fraction was passed through a DE-52 column and then fractionated with 20–30% ammonium sulphate. Precipitated proteins were resuspended, dialysed overnight, and purified further with a Mono-Q column and a Superdex-200 column (Amersham).

Crystallization and surface-entropy-reduction mutagenesis. All crystallization trials were performed with the hanging-drop vapour-diffusion method by mixing 1 µl of protein solution (12 mg ml^{−1}) with 1 µl of reservoir solution. EscJ(21–190) crystallizes under several conditions containing PEG3350, but none of these crystals diffracted X-rays. Mutations targeting different regions of EscJ were constructed with the Quikchange method (Stratagene). The mutant (E62A/K63A/E64A) yielded crystals suitable for structure determination. Complementation experiments showed that this triple mutation does not affect the function of EscJ (Supplementary Fig. 3). The mutant crystallizes in 0.2 M diammonium hydrogen phosphate, and the crystals grew to maximal size over 2–3 days at 18 °C. A heavy-atom derivative was prepared by soaking native EscJ crystals in 5 mM *p*-chloromercuribenzoate (PCMB) in reservoir solution for 32 h.

Data collection and structure determination. EscJ crystals were harvested by sequentially soaking in mother liquor containing 17.5% and 35% glycerol, and then flash-frozen to 100 K. All data were collected at beamline 8.2.1 of the Advanced Light Source (ALS). A redundant data set was collected from a PCMB-derivatized crystal near the mercury peak and a native crystal (Supplementary Table 1). Data were processed with Mosflm¹⁸ and scaled with SCALA¹⁹. Heavy-atom positions were found and refined with SOLVE²⁰, and a high-quality experimental map was obtained after density modification with RESOLVE²¹. The model was built manually with Xfit²². Refinements were performed with

CNS²³ and finally with Refmac⁵²⁴. The refined model is complete except for residues 92–97 and 134–140 in all four molecules of the asymmetric unit and the C-terminal five to six residues in three of four molecules of the asymmetric unit, which were disordered. The final model has no Ramachandran violations, with 96.2% of the dihedral angles in the most favoured regions. The triple mutation, which localizes to a solvent-accessible loop region in domain 1, does not affect the folding and organization of secondary structures (Supplementary Fig. 3).

Molecular modelling. The rotation and translation parameters of one monomer to the other three molecules in the EscJ tetramer were determined with LSQKAB²⁵. A 'flattened' tetramer was generated by setting the z-translation to zero, and a sixfold rotation was then applied. The resulting model was refined with CNS with manual inspection. The final model contains no unfavourable molecular contacts.

Illustrations. Surface calculations were performed with GRASP²⁶. All other figures were generated by Molscript²⁷, Raster3D²⁸, Xfit²² and Pymol (<http://pymol.sourceforge.net>).

Received 20 November 2004; accepted 14 March 2005.

- Ghosh, P. Process of protein transport by the type III secretion system. *Microbiol. Mol. Biol. Rev.* **68**, 771–795 (2004).
- Marlovits, T. C. *et al.* Structural insights into the assembly of the type III secretion needle complex. *Science* **306**, 1040–1042 (2004).
- Kubori, T. *et al.* Supramolecular structure of the *Salmonella typhimurium* type III protein secretion system. *Science* **280**, 602–605 (1998).
- Blocker, A. *et al.* Structure and composition of the *Shigella flexneri* 'needle complex', a part of its type III secretin. *Mol. Microbiol.* **39**, 652–663 (2001).
- Kimbrough, T. G. & Miller, S. I. Assembly of the type III secretion needle complex of *Salmonella typhimurium*. *Microbes Infect.* **4**, 75–82 (2002).
- Kimbrough, T. G. & Miller, S. I. Contribution of *Salmonella typhimurium* type III secretion components to needle complex formation. *Proc. Natl Acad. Sci. USA* **97**, 11008–11013 (2000).
- Galan, J. E. & Collmer, A. Type III secretion machines: bacterial devices for protein delivery into host cells. *Science* **284**, 1322–1328 (1999).
- Sukhan, A., Kubori, T., Wilson, J. & Galan, J. E. Genetic analysis of assembly of the *Salmonella enterica* serovar *Typhimurium* type III secretion-associated needle complex. *J. Bacteriol.* **183**, 1159–1167 (2001).
- Crago, A. M. & Koronakis, V. *Salmonella* InvG forms a ring-like multimer that requires the InvH lipoprotein for outer membrane localization. *Mol. Microbiol.* **30**, 47–56 (1998).
- Aizawa, S. I. Flagellar assembly in *Salmonella typhimurium*. *Mol. Microbiol.* **19**, 1–5 (1996).
- Burghout, P. *et al.* Structure and electrophysiological properties of the YscC secretin from the type III secretion system of *Yersinia enterocolitica*. *J. Bacteriol.* **186**, 4645–4654 (2004).
- Linderoth, N. A., Simon, M. N. & Russel, M. The filamentous phage pIV multimer visualized by scanning transmission electron microscopy. *Science* **278**, 1635–1638 (1997).
- Nouwen, N. *et al.* Secretin PulD: association with pilot PulS, structure, and ion-conducting channel formation. *Proc. Natl Acad. Sci. USA* **96**, 8173–8177 (1999).
- Jones, C. J., Macnab, R. M., Okino, H. & Aizawa, S. Stoichiometric analysis of the flagellar hook–(basal-body) complex of *Salmonella typhimurium*. *J. Mol. Biol.* **212**, 377–387 (1990).
- Sekiya, K. *et al.* Supermolecular structure of the enteropathogenic *Escherichia coli* type III secretion system and its direct interaction with the EspA-sheath-like structure. *Proc. Natl Acad. Sci. USA* **98**, 11638–11643 (2001).
- Suzuki, H., Yonekura, K. & Namba, K. Structure of the rotor of the bacterial flagellar motor revealed by electron cryomicroscopy and single-particle image analysis. *J. Mol. Biol.* **337**, 105–113 (2004).
- Thomas, J., Stafford, G. P. & Hughes, C. Docking of cytosolic chaperone-substrate complexes at the membrane ATPase during flagellar type III protein export. *Proc. Natl Acad. Sci. USA* **101**, 3945–3950 (2004).
- Leslie, A.G.W. Recent changes to the MOSFLM package for processing film and image plate data. *Joint CCP4+ESF-EAMCB News. Protein Crystallogr.* no. 26 (1992).
- Evans, P. R. Data reduction. *Proc. CCP4 Study Weekend on Data Collection and Processing* 114–122, (1993).
- Terwilliger, T. C. & Berendzen, J. Automated MAD and MIR structure solution. *Acta Crystallogr. D* **55**, 849–861 (1999).
- Terwilliger, T. C. Maximum-likelihood density modification. *Acta Crystallogr. D* **56**, 965–972 (2000).
- McRee, D. E. XtalView/Xfit—A versatile program for manipulating atomic coordinates and electron density. *J. Struct. Biol.* **125**, 156–165 (1999).
- Brunger, A. T. *et al.* Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr. D* **54**, 905–921 (1998).
- Murshudov, G. N. Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr. D* **53**, 240–255 (1997).
- Kabsch, W. A solution for the best rotation to relate two sets of vectors. *Acta Crystallogr. A* **32**, 922–923 (1976).
- Nicholls, A., Sharp, K. A. & Honig, B. Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons. *Proteins* **11**, 281–296 (1991).
- Kraulis, P. J. MOLSCRIPT: A program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallogr.* **24**, 946–950 (1991).
- Merritt, E. A. B. & Bacon, D. J. Raster3D: photorealistic Molecular Graphics. *Methods Enzymol.* **277**, 505–524 (1997).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank A. L. Lovering, C. P. C. Chiu and P. I. Lario for discussions; H. Law, K. Hayakawa, Y. Luo and Y. Wu for involvement in the early stages of the project; and the staff at the Advanced Light Source beamline 8.2.1 for data collection time and assistance. C.K.Y. is supported by fellowships from the Natural Sciences and Engineering Research Council of Canada and the Michael Smith Foundation for Health Research. N.C.J.S. and B.B.F. thank the Howard Hughes Medical Institute International Scholar Program, Canadian Institutes of Health Research and the Canadian Bacterial Diseases Network for funding. Funding for this project also came from grants from the NIH to S.I.M.

Author Contributions C.K.Y. completed the structural determination, analysis and modelling of EscJ, M.V. assisted in purification and crystallization of EscJ, R.A.P. developed the EscJ purification procedure, and E.A.F. did the original cloning of EscJ under the supervision of N.C.J.S. T.G.K. and H.B.F. performed the EM, labelling, and mass spectrometry experiments on *Salmonella* NCs under the supervision of S.I.M. and N.A.T. performed the EscJ localization and complementation assays under the supervision of B.B.F.

Author Information Coordinates and observed structure factors have been deposited to the Protein Data Bank under accession code 1YJ7. Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to N.C.J.S. (natalie@byron.biochem.ubc.ca).

-  FOCUS
-  SPOTLIGHT
-  RECRUITMENT
-  ANNOUNCEMENTS
-  EVENTS

naturejobs

Making a move on mobility

The European Union (EU) uses its Marie Curie Fellowship programme actively to promote mobility among young scientists. It does this by providing funding for graduate students and postdocs as long as they train in an EU country other than their own. Mobility seems like a good goal. In the best cases, it fosters the flow of ideas, builds collaborations, and sends money and talent into under-funded regions.

But mobility has its down sides, too. Phrases such as "brain drain" and "off-shoring" have significant negative connotations. At this stage, it is hard to know whether in the long term the EU's scheme will have a positive or negative outcome for the young researchers it has been funding. There are very few centralized data on the matter. Indeed, both the US National Postdoctoral Association and the National Academy of Sciences this year issued a plea for more international data.

That situation could be about to change thanks to the Young European Biotech Network (YEBN). Last month, at the World Life Sciences Forum BioVision 2005 in Lyon, France, the network launched a survey. This aims to identify

the key problems encountered by Europe's young researchers, scientists and technicians. The data collected will help European policy-makers to review and refine their strategies, by giving them a clearer picture of what effects their policies have had. And it will benefit young researchers, because they will know whether their own problems are personal or universal.

The success of this project comes down to numbers. The YEBN is aiming for 1,000 responses, but more would be better. Young European researchers who take the time to fill out the survey and ask friends and colleagues to do the same will go a long way towards answering the questions about the benefits and pitfalls of mobility, and bring to light other issues of concern for young researchers. The survey is online now at www.yebn.org.



Paul Smaglik, *Naturejobs* editor

CONTACTS

Publisher: Ben Crowe
Editor: Paul Smaglik
Marketing Manager: David Bowen

European Head Office, London
 The Macmillan Building, 4 Crinan Street
 London N1 9XW, UK
 Tel: +44 (0) 20 7843 4961
 Fax: +44 (0) 20 7843 4996
 e-mail: naturejobs@nature.com

Naturejobs Sales Director:
 Nevin Bayoumi (4978)
European Sales Manager:
 Andy Douglas (4975)

UK/RoW/Ireland:
 Nils Moeller (4953)
 Irene Viglia-Atton (4944)
Scandinavia/Spain/Portugal:
 Evelina Rubio Håkansson (4973)
Natureevents: Silke Opstrup (4994)
France/Switzerland:
 Amelie Pequignot (4974)

Advertising Production Manager:
 Billie Franklin
 To send materials use London
 address above.
 Tel: +44 (0) 20 7843 4814
 Fax: +44 (0) 20 7843 4996
 e-mail: naturejobs@nature.com

Naturejobs web development: Tom Hancock
Naturejobs online production: Niamh Shields

European Satellite Office
Germany/Austria/Italy/
The Netherlands/Belgium:
 Patrick Phelan
 Tel: +49 89 54 90 57 11
 Fax: +49 89 54 90 57 20
 e-mail: p.phelan@nature.com
 Reya Silao
 Tel: +44 (0) 20 7843 4970
 e-mail: r.silao@nature.com

US Head Office, New York
 345 Park Avenue South,

10th Floor, New York, NY 10010-1707
 Tel: +1 800 989 7718
 Fax: +1 800 989 7103
 e-mail: naturejobs@natureny.com

US Sales Manager: Peter Bless

Japan Head Office, Tokyo
 MG Ichigaya Building (5F),
 19-1 Haraikatamachi,
 Shinjuku-ku,
 Tokyo 162-0841
 Tel: +81 3 3267 8751
 Fax: +81 3 3267 8746
Asia-Pacific Sales Director: Rinoko Asami
 e-mail: rasami@naturejpn.com

Dial M for middleware

Dirty work down on the Pharm.

David Hall

I was in a lab again — the first time since I gave up trying to get a PhD in biochemistry and joined the police force. This one at PHunnyPharm had gleaming, state-of-the-art equipment. Did you ever use a fraction collector years ago? Big clunky thing that you left running overnight filled with battered old test tubes? You came in the next morning to discover that it had flung precisely the tubes of interest onto the floor. This lab's machine had a neat continuous spiral of plastic tubes, all still working away. The floor was a real mess, though. Even in my day we drew the line at dead bodies.

Judy, the technician who'd found him, told me that he was the company's star researcher, Plantagenet Benfield. He'd been battered to death sometime between midnight and three in the morning, according to the pathologist.

PHunnyPharm's proprietor, Simon Clark, was taking his time to arrive, so I talked to Richard in Security. He was obviously glad of the chance to show off his gee-whiz electronic gadgets.

The roar of a sports car heralded Clark's arrival. I had liked Benfield better. He hadn't looked so smug — dead.

"Simon, did Benfield often work at night?"

"He shouldn't have done. All alone. Health and safety issues. But he was very dedicated. Soldiering away on his own."

"Do you know what he was working on?"

"It'll be in his notebooks."

Twice he had felt the need to assure me that Benfield had been working alone. I went back to Judy so she could talk me through the equipment she'd set up for him. The extract went in here; a set of columns fractionated it into its component parts here; which were sent as a stream to that lovely fraction collector, there. Samples of the fractions were taken, sent round and round in tubes, and came out there.

"What happens here?"

"The samples are bioassayed for psychedelia. All of this is under computer control through a middleware system that allows different units to talk to each other. Plantagenet used it to program a series of tests on selected fractions. For instance, if the machine finds a frac-

tion containing a psychedelic, it then can also have it assayed for, say, blood-vessel constriction — some psychedelics act against cluster headaches — by this box, or for something else by that box. If it found anything interesting it would call his mobile."

"For a chat?"

"Plantagenet had different ringtones for calls from different parts of the set-up. A call from the bioassay made it play *The Yellow Rose of Texas*. Details would be in a text."

"Do you know why Benfield was working late last night?"

"He wasn't planning to."

The scene-of-crime officer had given me Benfield's mobile. I looked up his last incoming call. Midnight. I showed the number to Judy.

"Does this mean anything to you?"

"It's the middleware direct line."

"The system called him in then. What exactly had he been working on?"

"The same as other people here. Magic

mushrooms — they have a research licence. I think Plantagenet was looking for psychedelics in some of the company's extracts. It had upset Simon."

Clark had implied that he didn't know what Benfield had been working on.

"Upset him how?"

"Our patented extracts are psychedelic-free, so they're legal. If Plantagenet had found psychedelics in them, it would have been death — commercial death. An invalid patent."

Time to talk to the boss again.

"So what do you think happened last night, Simon?"

"We hold a lot of psychedelic drugs here. Benfield was disturbed by an intruder. Bam! No great mystery, I think. Unfortunately the intruder must have taken the hard disk. There are no CCTV records for last night."

Dee dum dee dee dee dum dee

"Is that your mobile, Simon?"

"Excuse the tune. Someone's idea of a joke." He put the phone quickly to his ear. "OK. Sorry, but you know..."

I pulled the phone straight from his hand — there was a text message, but silence at the other end.

"I do know *The Yellow Rose of Texas*. I think Benfield's equipment has just tried to call him — but it also called you. That happened last night, didn't it? Benfield's analyser found something and called him. You'd been suspicious, and got his equipment to call you as well. You came in..."

"You can't prove that!"

"Because you're not on film? You ought to take more interest in your own security systems. Your fence has microphones all over it. I have compared the sound of a car engine it recorded last night, with that of your car engine from this morning. They're the same."

"You couldn't have done all that since you got here."

I waved my mobile at him. "The police use middleware too. You're nicked."

A good morning's work. Especially as I was now the only one who knew which of their compounds was a potential headache cure. Clark should have read that last text.

David Hall used to be a biochemist, but got a life, and then a wife, great niece of a celebrated mystery writer, who would have relished all of today's possibilities. He is aware that nostalgia isn't what it used to be.



JACEY